ERIC Document Resume | See Instruction Sheet | DATE OF RESUME 8/11/65 | DATE RECEIVED

A. PUBLICATION DATA

1. ERIC NUMBER ED000451
2. ORIGINATOR CONTROL NO(S). 7203
3. IDENTIFICATION NO. N.A.

4. AUTHOR(S): Stuit, Dewey B., and others

5. TITLE OF DOCUMENT ABSTRACTED:

Survey of College Evaluation Methods and Needs

6. PUBLICATION TITLE: N.A.

7. EDITOR(S): N.A.

8. PLACE OF ORIGIN
ORG. Educational
ST. Testing Service
CITY, STATE Princeton, N.J.
TELE.
EXT.
INST/LOC CODE

9. PUBLISHER OR SOURCE: Educational Testing Service

10. VOL.	11. PAGES	12. MO./YR.	13. TOTAL PP.
N.A.	N.A.	56	225

DO NOT WRITE IN SPACE BELOW

MICROFILM COPY AVAILABLE
YES ☒ NO ☐

COPYRIGHT GRANTED
YES ☐ NO ☒

COPYRIGHT RELEASE FILED
YES ☐ NO ☐

TYPE OF REPORT
PROPOSAL ☐
IN-PROGRESS ☐
FINAL ☒
FOLLOW-UP ☐

OTHER ERIC REFERENCES:

ERIC SATELLITE CODE

B. INDEXING DATA

14. SUBJECT AREA

15. TYPE OF ENDEAVOR: survey

16. INDEXING TERMS (6-12)

		teaching methods		
surveys	objectives	students		
evaluation	curriculum	student services		
measurement	faculty	effectiveness		

17. ABSTRACT (MAX. 250 WORDS)

Survey of problems of college evaluation. Identifies areas where additional and improved methods, instruments, and standards are needed most. Methods and materials for evaluating the following aspects of a college were surveyed, and the results reported: Institutional objectives, curriculum, faculty, instructional effectiveness, student body, and student personnel services.

ERIC No.

18. SOURCE OF ABSTRACT

C. FUNDING DATA

19. SUPPORTING ORGANIZATION

20. DIVISION OR DEPT.

21. CONTRACT NO.

PROJECT NO.

22. TOTAL COST OF PROJECT

23. DURATION OR PROJECT

U.S. Department of Health, Education, and Welfare
Office of Education

ERIC Form 1

~~IV-20~~
IV-17

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

7203

"SURVEY OF COLLEGE EVALUATION METHODS AND NEEDS,"

A Report to the Carnegie Corporation

Dewey B. Stuit
Gerald C. Helmstadter and
Norman Frederiksen.

Educational Testing Service
Princeton, New Jersey
December, 1956

IV-20
ED000451

SURVEY OF COLLEGE EVALUATION METHODS AND NEEDS

A Report to the Carnegie Corporation

Dewey B. Stuit
Gerald C. Helmstadter
Norman Frederiksen

Educational Testing Service
Princeton, New Jersey
December, 1956

Acknowledgments

This report was reviewed by Dean Russell M. Cooper, Professor T. R. McConnell, and Professor C. Robert Pace. A portion of the manuscript was reviewed by Professor Robert L. Ebel. The authors are grateful to these reviewers for many helpful suggestions. However, any errors of fact or reasoning in this report are the sole responsibility of the authors.

TABLE OF CONTENTS

CHAPTER I. PURPOSE OF THE STUDY 1

CHAPTER II. INSTITUTIONAL OBJECTIVES 5

Introductory Statement 5

Evaluation Problems 6

Appropriateness of the objectives 6

Precision of the statement of objectives 8

Adequacy of the process of formulating objectives 12

Extent to which objectives have been elaborated 13

Faculty and administration concern about objectives 14

Techniques of Evaluation 15

Study and discussion 15

Evaluation by opinions 16

Follow-up studies 18

Comparison of faculty conceptions of objectives with published statements 20

Evaluation Needs 22

Refinement of the judgment process 22

Improved methods for determining the institution's role 25

Knowledge of what kinds of discussions tend to lead to adequate statements of objectives 27

Better correlation between student and institutional objectives 28

Summary of Evaluation Needs 29

Summary Table for Chapter II 31

CHAPTER III. THE CURRICULUM 33

Introductory Statement 33

Evaluation Problems 33

Implementation of the institution's objectives 33

Recognition of the instructional needs of students 34

Recognition of individual differences among students 35

Articulation with curricula of other educational units 35

Meeting the needs of society 36

Techniques of Evaluation 36

Evaluation by opinions 37

Study and discussion 38

Follow-up studies 39

Comparison with other institutions 42

Educational and psychological tests 43

Persistence studies 45

Measurement of duplication 46

Evaluation by experimentation 47

Evaluation Needs 49

Better use of opinions or judgment 49

Better experimental evaluations 50

Better information about curricular practices 53

Better follow-up studies 53

More and better studies of general vs. specialized courses 55

Summary of Evaluation Needs 56

Summary Table for Chapter III 61

CHAPTER IV. THE FACULTY 63
Introductory Statement 63
Evaluation Problems 64
    Qualifications of the faculty 64
    Scholarly production 66
    Attitudes toward scholarship, teaching, and students 68
    Personal attributes in relationship to colleagues 68
    Observance of professional ethics 70
    Morale of the faculty 71
    Contribution to school and community activities 73
Techniques of Evaluation 74
    Biographical data sheets 74
    Rating scales 76
    Scholastic records and test scores 79
    Letters of recommendation and personal interviews 80
    Attitude scales 81
    Amount of turnover 82
Evaluation Needs 83
    Better evaluation of quality of preparation 83
    Better ratings of faculty qualifications 84
    Better norms for evaluation of biographical data 86
    Better assessment of attitudes 87
    Better evaluation of personal attributes 88
    Evaluation of the faculty as members of a group 89
    Improved letters of recommendation and interviews 90
    Monitoring of the teaching process 91

CHAPTER IV. THE FACULTY (continued)

Summary of Evaluation Needs 92
Summary Table for Chapter IV 97

CHAPTER V. INSTRUCTIONAL EFFECTIVENESS 99
Introductory Statement 99
Evaluation Problems 99
Measurement of student performance 99
Measurement of student satisfaction 101
Measurement of supervisor satisfaction 102
Techniques of Evaluation 103
Evaluation by opinions 104
Systematic observation of teaching materials and processes 106
Teacher-made examinations 107
Standardized achievement examinations 109
The pre-test and post-test method 111
Attitudes and appreciations tests and scales 113
Follow-up studies 117
Evaluation Needs 120
Better measures of achievement built around particular instructional objectives 121
Better test norms 124
Better techniques for assessing changes in attitudes and appreciations 125
Better measures of critical thinking and problem-solving 126
Better methods of observation and rating 127
Better definitions of good teaching 129
Summary of Evaluation Needs 130
Summary Table for Chapter V 133

CHAPTER VI. THE STUDENT BODY 135

Introductory Statement 135
Evaluation Problems 135
Size 135
Quality 137
Achievement of the student body 138
Morale of the students 139
Personal and demographical characteristics 139
Techniques of Evaluation 140
Tabulation of factual information 140
Personal data sheet 141
Aptitude and achievement tests 143
Questionnaires and attitude scales 144
Evaluation Needs 146
Studies of school characteristics 146
Development of appropriate measuring devices 148
Studies of the relationship among measures 151
Gathering of normative data 151
Summary of Evaluation Needs 154
Summary Table for Chapter VI 157
CHAPTER VII. STUDENT PERSONNEL SERVICES 159
Introductory Statement 159
Evaluation Problems 160
Extent of the services provided 160
Articulation with the academic program 160
Qualifications of the personnel 161
Student performance 162
Student and faculty satisfaction 163
Satisfaction of the supervisor or outside expert 163

CHAPTER VII. STUDENT PERSONNEL SERVICES (continued)

Techniques of Evaluation 164
    Check lists 164
    Survey of factual information 165
    Evaluation by opinions 165
    Ratings 166
    Achievement measures 167
    Experimentation 169
Evaluation Needs 171
    Studies of procedures 171
    Inter-school norms 172
    Needs analysis 173
    Adjustment measures 174
    Comparisons of evaluation techniques 175
Summary of Evaluation Needs 175
Summary Table for Chapter VII 177

CHAPTER VIII. COMPREHENSIVE EVALUATIONS 179
Introductory Statement 179
Evaluation Problems 179
    Performance of graduates 179
    Scholarly production of the faculty 180
    Scholarly atmosphere 181
    Meeting the needs of society 182
    Impact on higher education 183
Techniques of Evaluation 184
    Ratings by experts 184
    Follow-up studies of graduates 186
    Achievement measures 193

CHAPTER VIII. COMPREHENSIVE EVALUATIONS (continued)

Evaluation Needs 194

Better rating methods 194

Better follow-up studies 195

Better achievement measures 196

Summary of Evaluation Needs 196

Summary Table for Chapter VIII 199

CHAPTER IX. RECOMMENDATIONS 201

Direct Evaluation Needs 202

Development of adequate tests 202

Useful interpretive data 204

Improvement in the process of group judgment 209

Need for Background Information 211

What constitutes good practice? 211

What abilities are needed? 212

Summary 214

REFERENCES 217

Chapter I

PURPOSE OF THE STUDY

The college or university, no less than the business or advertising agency, must concern itself continually with evaluation of its own operations. It must examine critically all facets of its functioning in order to determine where modifications of curricula, teaching practices, or student services are needed. Accreditation by state and regional associations has done much to raise standards and improve educational practices, but in the long run it is the evaluation of an institution by itself which will be most productive of growth toward the full attainment of its objectives.

In order for self-evaluations to be performed effectively, it is of course necessary to use suitable methods and tools and to have appropriate standards against which to judge performance. Lack of suitable tools and standards in many areas has undoubtedly seriously hindered self-evaluation efforts by colleges. The purpose of the present study is to survey the problems of college evaluation with a view to identifying those areas where additional and improved methods, instruments, and standards are most needed.

Methods and materials for evaluating the following aspects of a college have been surveyed, and the findings are reported in later chapters of this report:

1. Institutional objectives
2. Curriculum
3. Faculty
4. Instructional effectiveness
5. Student body
6. Student personnel services

In addition, methods which are applicable to an over-all evaluation of the institution were investigated.

In each chapter, three main divisions are addressed to answering the following three kinds of questions:

(1) What are the principal dimensions or aspects of the evaluation problem under consideration in the chapter?

(2) What methods and materials are currently available for making evaluations of these dimensions?

(3) What methods and materials are needed in order to improve these evaluations?

A final chapter summarizes the evaluation needs that have emerged from the detailed consideration of the problems and presents them in the form of broad areas of research which need to be done.

In using this report, it would be well for the reader to keep in mind that our concern is the evaluation of the institution, or of some subdivision of the institution, and not evaluation of individual students or faculty members. This distinction is important because of considerations having to do with reliability of measurement. Tests which are inadequate for the fair evaluation of an individual may be quite satisfactory when used to provide data for an entire group. To be sure, the evaluation of particular facets of an institution, e.g., faculty, students, and quality of instruction, requires that one obtain information about individual faculty members and students. However, in interpreting the data we are concerned with the data for groups -- means, medians, standard deviations and correlation coefficients -- not with the specific standing or performance of Professor A or of student X. In the final analysis, the performance of individual faculty members and students determines what the group performance will be, but for our purposes it is collective performance which is of primary concern.

It may also be well to make clear several other limitations with respect to the aims of this survey. We are not concerned with reviewing all the studies which have been made in the field of college evaluation. When studies are cited, it is not for the purpose of reporting results but rather to illustrate or describe techniques which appear to be useful in college evaluation. We are not concerned with the technical details of various measurement techniques, such as psychological tests, personality inventories, biographical data sheets, or attitude scales. It should also be made clear that this report does not aim to present a blueprint on how to improve a college or even how to conduct a college evaluation study. It is aimed solely at surveying the evaluation methods and interpretive data which are available and at pointing out the techniques and interpretive data which are needed in order for a college to make better self-evaluations.

It is of course impossible to make evaluations without facing the problem of defining goodness. Definitions of goodness will undoubtedly vary from college to college, because of such considerations as size, type, location, and purpose of the institution. While the discussions of evaluation methods which follow undoubtedly are based on certain assumptions of the authors about what is good and bad about a college, it is our intent to leave to each college the problem of defining goodness in the light of its own characteristics. The evaluation methods discussed are, generally speaking, concerned with how to get data which are thought to be potentially useful in college evaluation. How the data are to be used in a specific case is a matter to be decided by those responsible for the institution concerned.

PRECEDING PAGE BLANK- NOT FILMED

Chapter II

INSTITUTIONAL OBJECTIVES

Introductory Statement

The total educational program of any institution of higher education should be determined by the objectives which the institution is seeking to achieve. The breadth and depth of the curriculum, the extensiveness of its student personnel services, the facilities for research and service -- all these should ultimately be determined by the educational goals which the institution has established for itself. Nearly all college catalogs contain some statement of institutional objectives, generally phrased in rather impressive language. The important question is whether the institution has a realistic set of objectives and a program of activities designed to achieve them in practice. How can an institution evaluate its stated objectives and determine their validity? How can it determine whether it has a program of activities which will enable it to achieve its stated objectives? The study of these questions is easily slighted, but most educators would agree that every college and university in the country could well give them frequent, thoughtful consideration. This chapter is concerned with the first question; subsequent chapters will be concerned with the second question.

The North Central Association of Colleges and Secondary Schools in its accrediting procedures has emphasized that each institution should determine carefully the particular educational objectives which it is best equipped to serve. This means, in effect, that each institution must exercise choice in defining and listing its objectives. In the words of Haggerty (49 , p. 7):

> "No institution has, or ever will have, sufficient resources to serve all the purposes of higher education. Nor does America require a multiplication of higher institutions all attempting the same thing. What it does need is a total pattern of institutions in which each, by serving its limited functions well, will help to create a system of higher education serving all students according to their individual needs -- a system in which society will find all the educational services that are useful to it. In such a system each college must select its particular functions, determine the direction of its resources, limit its activities, frankly and publicly declare its intentions, and adhere to them in practice. By so doing, a socially needed institution will carve out for itself a unique place among the whole group of higher institutions."

The determination of an institution's objectives is extremely important and far-reaching in its implications. What, then, are the characteristics or attributes of a desirable set of objectives? What techniques can be used to determine the adequacy of an institution's stated objectives?

Evaluation Problems

Appropriateness of the objectives. The first, most crucial dimension of the objectives-evaluation problem is that of determining the appropriateness of an institution's objectives -- for the students taught, the community directly served, society at large, and for the institution's facilities, including physical plant, faculty, and geographical location.

Colleges differ considerably in the types of students which they enroll. Some institutions, such as engineering colleges, for example, serve primarily those planning to enter a particular professional group; others serve a student clientele planning to enter a very wide variety of occupational fields. Some institutions, such as state universities, admit students with considerable variation in intellectual abilities; others matriculate only those who are in the very top stratum of academic aptitude and achievement. Our society needs a rich variety of colleges and universities with distinct differences in educational goals. The important point is that the objectives of a particular school should accord with the background, interests, and ambitions of its particular student body. No factor seems more important to the determination of the appropriateness of an institution's objectives than the nature of the student body which it has enrolled or is seeking to enroll.

A second facet of the "appropriateness" dimension is that of the adequacy of the objectives to the community or communities directly served. The "community" may be a city, state, region, denomination, or other definable group. Within our democratic scheme of things it is not only appropriate but also desirable that a college or university should on occasions aim to serve particular communities. If an institution is to serve a given community, then the institution's objectives should give evidence of this fact. For example, a college or university which is supposed to serve the adult as well as the late-adolescent citizen must show in its statement of objectives that it recognizes this particular mission. The extent to which this community responsibility is recognized becomes one criterion by means of which the institution's objectives can be evaluated.*

Important as the local community may be, every college must also ask itself whether it is serving the needs of society as a whole. All institutions of college grade must, for example, recognize that their students will soon be adult citizens in a democratic society. Even if the particular function of the school may be highly vocational or professional, it cannot ignore the fact that no adult spends all of his time in vocational pursuits. The extent to which the institution recognizes this fact is another aspect of the problem of measuring the adequacy of its objectives.

---

*Certain junior colleges which have adopted the community college philosophy illustrate very well how an institution can develop in response to community needs. (3 , pp. 94-117.)

Colleges and universities, like individual citizens, sometimes "bite off more than they can chew." It is not uncommon for this to be done in order to attract larger student bodies or to enlist a broader base of support. Limitations of faculty, physical plant, and geographical location may stand in the way of the realization of published objectives. Undergraduate institutions are sometimes tempted to undertake professional training programs which they are not, in staff or in physical facilities, equipped to handle. Junior colleges may aspire to be four-year colleges, and four-year colleges sometimes aspire to add a graduate program, when they are not really equipped to function at this level. To be ambitious is a laudable tendency, but it may result in the institution's over-reaching itself. The objectives of an institution and its facilities must be in harmony with one another; the magnitude of this "harmony" is one aspect of the appropriateness of its objectives.*

Precision of the statement of objectives. It is the rule, not the exception, to find that a college's objectives are stated in general rather than specific terms. As mentioned in the previous section, the writers of college catalogs are prone to prepare eloquent statements of objectives which may have little direct bearing upon the construction of the curriculum or upon the teaching process. The same can be said of many general speeches and articles written by college administrators for the general public.

---

*A recent attempt to formulate differential functions for three groups of institutions -- junior colleges, state colleges, and universities -- has been made for the state of California. (85 , pp. 49-220.)

The assumption that it is necessary to have imprecise statements of objectives for public relations purposes is questionable. Rather, it would seem that objectives should be so stated that their meaning is clear to prospective students and their parents and to the college faculty and administration who must work to attain the objectives.

To indicate that the college should develop the student intellectually, physically, emotionally, aesthetically, personally, and spiritually is, perhaps, satisfactory as a general descriptive statement, but not very meaningful as a guide to the planning of the educational program. Instead of saying simply that the student should be developed intellectually, it would be better to make a more specific statement such as: (1) he should develop his reading, writing, speaking, and calculating skills to meet the demands of college courses and the responsibilities of citizenship, (2) he should study the leading ideas, principles, and concepts in each of the major divisions of knowledge, and (3) he should acquire especial competence in some particular area of knowledge. These statements are also quite general, but at least one can see a relationship between the described objectives and the behaviors which the college hopes to promote in its students. In brief, the clarity with which objectives are stated and the degree of their correlation with actual behavior would seem to be an important dimension, the measurement of which should help considerably in determining the quality of the objectives.

One of the most interesting recent attempts at achieving precision in the statement of educational objectives and illustrating how precision of communication can be achieved has been prepared by a group of test constructors, curriculum directors, and teachers (7). The classification system developed by this group of educators and examiners is modelled after library classification systems. It was recognized that, ideally, all educational objectives which can be stated in behavioral form should find a place in the taxonomy, including the cognitive, affective, and manipulative or motor skills areas. It was decided, however, to concentrate attention first on the cognitive area, including knowledge, intellectual abilities, and intellectual skills. The main headings in the system are:

1.00 Knowledge
    1.10 Knowledge of Specifics
    1.11 Knowledge of Terminology
    1.12 Knowledge of Specific Facts

1.20 Knowledge of Ways and Means of Dealing with Specifics
    1.21 Knowledge of Conventions
    1.22 Knowledge of Trends and Sequences
    1.23 Knowledge of Classifications and Categories
    1.24 Knowledge of Criteria
    1.25 Knowledge of Methodology

1.30 Knowledge of the Universals and Abstractions in a Field
    1.31 Knowledge of Principles and Generalizations
    1.32 Knowledge of Theories and Structures

2.00 Comprehension
    2.10 Translation
    2.20 Interpretation
    2.30 Extrapolation

3.00 Application

4.00 Analysis
    4.10 Analysis of Elements
    4.20 Analysis of Relationships
    4.30 Analysis of Organizational Principles

5.00 Synthesis
    5.10 Production of a Unique Communication
    5.20 Production of a Plan, or Proposed Set of Operations
    5.30 Derivation of a Set of Abstract Relations

6.00 Evaluation
    6.10 Judgments in Terms of Internal Evidence
    6.20 Judgments in Terms of External Criteria

The classification system proposed by Bloom and the Committee of College and University Examiners is obviously an aid to communication and a means of laying the groundwork for the building of examinations. The system places the emphasis upon clarity of definition, particularly upon the defining of objectives in terms of student behaviors. If applied to specific courses (e.g., college algebra) in a number of different institutions, the system could result in a more uniform understanding of what each course should contribute to the total educational program and of what should be measured in achievement tests. While not in itself a statement of objectives, it is a good illustration of what is meant by clarity and precision in the statement of educational objectives and might serve as a standard to which college statements of objectives might be compared for purposes of evaluation.

Another illustration of an attempt to state educational objectives in precise, behavioral terms is to be found in a study concerned with general education for members of the armed forces (80). The procedure in this instance was for a committee first to list ten objectives of general education. For each of these ten objectives, statements were prepared indicating the "behaviors" which the student should acquire in the attainment of each of the objectives, the outline used being, in each instance, (A) Knowledge and Understanding, (B) Skills and Abilities, (C) Attitudes and Appreciations. Following the formulation of the objectives and preparation of the statements of behaviors to be acquired, faculty committees prepared outlines of the courses which could be taught in the attainment of the objectives. While this particular study deals with general education, it illustrates very well how a clear statement of objectives can lead to the development of actual course materials, including textbooks, references, and teaching aids. Obviously, such a statement of objectives is also easier to evaluate than one expressed in more general terms.

Adequacy of the process of formulating objectives. If the sources of the statements of objectives of most institutions were tabulated, it would be found that they are largely the product of individual or committee thinking; in only a limited number of cases are they based upon data collected in systematic studies. The objectives and the curriculum of an institution such as Stephens College, Columbia, Missouri, are a notable exception.

Educators generally would agree that any statement of objectives must ultimately be made on the basis of experienced and discerning judgment. There is no gainsaying the importance of a philosophical framework for educational objectives. Institutions exemplifying a particular denominational point of view may have a highly unified and pervasive philosophical orientation. Others, less single-minded in philosophical attitude, may find it possible to center on certain behavioral outcomes -- intellectual, moral, and emotional -- congruent with more than one fundamental position. Such an analysis has actually been made for liberal education by Smith (65).

All those connected with higher education in America would agree that our objectives in higher education must be based upon a democratic philosophy of government, of economics, and of society.

And yet, while granting the validity and power of the ideas and judgments of the human mind, it would seem that wherever possible this judgment and imagination should be based upon verifiable facts. If, for example, we say that colleges and universities should prepare young people for life in a democracy, we should know as definitely as we can what are the problems and responsibilities of the adult in a democratic society. If we are to prepare young people for a particular professional field, we should know as much as we can about the knowledge and skill required for the successful practice of that profession. The college or university which has gone to the "laboratory of life" to seek facts which may be of help in formulating objectives would seem to be proceeding on solid ground. In other words, the thoroughness or adequacy of the process by means of which objectives are defined would seem to be one means of determining the quality of these objectives.

Extent to which objectives have been elaborated. It is usually easier to speak in glittering generalities than to be concrete and specific. This is indeed true of statements of objectives! The disdain which some faculty members have for statements of objectives stems directly from the fact that in many cases these general objectives have not been elaborated sufficiently to make them effective in curriculum construction and daily teaching.

Any evaluation of objectives should, therefore, consider how these objectives have been stated, how much they are used in building the curriculum and in daily teaching. It is well and good for colleges to say that they are interested in the "whole" student; it is another thing to translate this philosophy into courses of study, counseling, advising, and extra-curricular activities. One can be quite certain that if objectives are stated in vague, general language, they will have little effect on the educational practice of the institution. If, on the other hand, the general statements of objectives have been elaborated into specific statements of behavior, and if these in turn are translated into educational activities, one can be reasonably sure that the institution's objectives are functional -- that they really influence the program of the institution and the education of its students.

Faculty and administration concern about objectives. The first step in the process of evaluating objectives is to sensitize the faculty and administration to the need for thinking about the subject. Unless there is a feeling on the campus that objectives should be studied or redefined, little serious thinking will be done about defining the role of the institution or ascertaining the extent to which it meets the needs of society. It is probably true that some administrative officers and standing committees on nearly every campus mention from time to time the necessity for re-examining the institution's objectives. Very frequently, however, this is mere superficial discussion in administrative offices or committee rooms. If objectives are to undergo a thorough evaluation, the entire faculty must be brought into the discussions and must feel a personal responsibility for the success of the undertaking. The extent to which the study of objectives becomes a matter of concern to the whole staff and the extent to which time and effort are devoted to the project would seem to constitute a significant dimension for the evaluation of objectives. The alert, progressive institution will be concerned about its objectives; the hidebound, unprogressive school will assume that its objectives, and their translation into educational programs, are adequate from generation to generation.

Techniques of Evaluation

Study and discussion. The oldest and most widely used method of evaluating educational objectives is that of study and discussion by individuals and groups. The charters or legislative acts establishing colleges and universities always contain a general statement of the objectives which the institution is supposed to achieve. One can guess that some of these were based upon very little study and discussion; in other cases there may have been a good deal of previous study and discussion. The latter was true, for example, in the case of the General College of the University of Minnesota (18). The study and discussion method, if based solely upon the experiences and observations of those participating in the evaluation process, may have many shortcomings. If based upon the best obtainable facts, opinions, and predictions, it becomes, in reality, the final step in the formulation of new objectives, the revision of old ones, or the placing of the stamp of approval on the existing list of objectives. This is as it should be, particularly if study and discussion are undergirded and supported by the best obtainable facts and analyzed by the most powerful and precise methods of educational research.

A new and interesting technique for securing some of the essential facts which may be needed in the evaluation of objectives has recently been proposed by Flanagan (20). As a first step Flanagan suggested that a tabulation be made of the adult activities for which colleges and universities attempt to prepare their students. The next step would be to identify the "critical requirements" for each of these activities, a critical requirement being defined as "an aptitude, ability, attitude, or other requirement for a particular activity which is crucial in the sense that it is frequently a factor causing either successful or unsuccessful participation." Flanagan's primary assumption is that these critical requirements should constitute a basis for the actual determination of objectives because they are based upon observed behavior, the observations being made by qualified observers and reported with a high degree of accuracy.

The technique proposed by Flanagan has been used primarily to determine the critical requirements for successful performance in various types of jobs. In this area it seems to have been quite successful. While it would appear that the procedure has interesting possibilities in the formulation (and, indirectly, the evaluation) of objectives, it should be kept clearly in mind that it can only provide some of the facts needed for study and discussion. Standing by itself it cannot formulate or evaluate an institution's objectives.

Evaluation by opinions. As suggested in the previous section, the evaluation of objectives can be achieved at least in part by collecting opinions as to the quality of an institution's objectives and subjecting these opinions to systematic, critical analysis. The persons in the best position to offer opinions are community leaders, faculty members, students, parents, and teams of experts. Opinions may be gathered by the use of questionnaires, check lists, rating scales, and interviews. All these techniques are simply means for obtaining opinions in objective form so they can be more easily tabulated, summarized, and analyzed. These opinions, however gathered, are verbalized statements of the judgments and feelings of individuals, based upon their observations and experiences. Since opinions are personal and subjective, they are subject to bias, prejudice, and provincialism. The task in collecting opinions is to give very careful directions and to collect opinions which are based upon facts. The evaluation of objectives by means of opinions requires that the human mind serve as a measuring instrument. The results obtained will be as good, or as bad, as the reliability and validity of that instrument.

One example of evaluating objectives by the systematic gathering of opinions is described by Eckert (18) in her account of what was done in the General College of the University of Minnesota. First, all members of the regular faculty (plus some others) were invited to write a statement of the objectives of the General College. These ideas were to be expressed in terms of the thinking and behavior of the students and were to emphasize those abilities, understandings, and attitudes which the faculty hoped a young person might possess at the end of his General College career. A total of 40 statements (from 40 faculty members) was received. A preliminary classification scheme was adopted as follows:

I. Knowledge as insight and understanding
II. Skills and abilities
III. Attitudes, interests, and appreciations
IV. Character of the individual's world view (philosophy)

The material presented by the faculty members was classified under the four headings, and all were then asked to read the statements appearing under each heading and to rate them as follows: Strongly agree, agree, no expressed attitude, goal questioned, disagree. Replies were received from 30 persons of the original 45. A committee consisting of two teaching members, a counselor, the director of curriculum revision, the director of the College, and the evaluator then reviewed the whole body of material and prepared a condensed statement of objectives for the College. Later each of the divisions and the instructors of individual courses prepared detailed statements of objectives for their own areas and related them to the College as a whole. In this manner, the stated objectives of the College were revised.

The evaluation of objectives by means of opinions can be no better than the experience, training, and judgment of the persons offering opinions will permit. Critics of the method have said that the collecting of opinions often results in little more than pooled ignorance. If the experience and judgment of the opinion givers are limited, the criticism would seem to be merited. On the other hand, if opinions are given by persons who are really qualified to give them, one can hardly think of a more appropriate method for the evaluation of objectives. The quality of those giving opinions and the care with which the opinions are analyzed would seem to be the key factors in determining the validity of this method of evaluation.

Follow-up studies. An important test of an institution's objectives is that of determining how far these objectives are realized in the lives of students. Obviously the objectives of a newly founded college cannot be studied by a follow-up study of its alumni; but eventually there comes a time when there are sufficient alumni or former students to make such a study possible. And if an entirely new program is to be undertaken it may prove feasible to study the alumni of other institutions of higher education. This is the procedure which was followed in the case of Stephens College, Columbia, Missouri (34 , p. 8).

The present curriculum of Stephens College is based upon the conviction that the purpose of general education is to serve the practical needs of its students (in this case college women), fitting them for a successful and satisfactory life as individuals and as members of society. In the view of Stephens College, this belief dictates that some institutions should concern themselves with the development of individual attitudes and abilities necessary to complete living -- not with the training of scholars in the strictly academic sense. With this philosophy as a basis, the college set itself the task of revising its program and objectives by making a follow-up study of college alumnae.

In order to discover the actual needs of students, a study to determine the nature of the adult activities of and the problems encountered by 300 college graduates was undertaken under the leadership of W. W. Charters. The sample was carefully drawn and the 300 women were then asked to keep accurate diaries for seven days together with comments concerning reasons for the activity or circumstances related to it. After the 300 diaries were returned, the first step was the identification of separate and distinct items. This produced a file of over 7500 cards, many bearing a sentence taken directly from a diary. These cards were then classified under 24 major headings. Next, a core curriculum was developed to teach the knowledge, skills, attitudes, and points of view which would enable students to meet the problems and responsibilities, as defined by these headings, encountered by women in modern society. The areas of instruction first selected were the following: Communications, social problems, physical health, mental health, aesthetic appreciation, consumer education, and philosophy of life. Later the following divisions were added: natural science, home and family relationships, and several vocational courses on an elective basis. The result of this whole study was a revision of the educational objectives of Stephens College and a complete reorientation of its program. The general approach employed could be applied with profit by many other colleges and universities.

The rating of the importance of various educational objectives in general education, and the extent to which alumni believe that their college experience helped them realize these objectives, is illustrated in a study by Pace (51) which was a part of a comprehensive self-study conducted by Syracuse University. A total of eighteen educational objectives was listed in a questionnaire sent to 5000 Syracuse alumni (2500 returns) with the requests (1) that the alumni rate on a five-point scale the degree of importance of each objective as a goal of a college education and (2) that they rate on a three-point scale the extent to which their college experience helped them to achieve these objectives. A somewhat similar study was conducted earlier in Minnesota, involving students from the University of Minnesota and certain other colleges in the state (1 , pp. 41-51). While the results of the Syracuse and Minnesota studies do not seem to reveal much which is especially new, studies of this kind do present a statement of how a group of students and alumni feel about their college experience. As faculty and administrative committees wrestle with problems of evaluating objectives, it may at least be reassuring to know that students and alumni agree with their thinking, and at best be helpful to have the benefits of new ideas or slants which students and alumni may be able to provide.

One of the most critical problems encountered in making a follow-up study is the adequacy of the sample studied. A closely related problem, of course, is the adequacy of the questions asked, or as in the case of the Stephens College study, the tasks to be performed -- in this instance keeping a diary for seven days. If a representative sample can be obtained, if the questions or the tasks really cover the subject, and if the results are carefully analyzed, a follow-up study can contribute a great deal to the evaluation and redefinition of an institution's objectives. Lacking these qualities, a follow-up study may not only fail to contribute positively to the reformulation of objectives but may actually give misleading results.

Comparison of faculty conceptions of objectives with published statements. It has already been mentioned that statements of objectives as they appear in college catalogs may not be highly correlated with actual practices in the educational program. Since classroom teaching constitutes so large and important a part of the total educational program, it would seem that one way to evaluate published objectives would be to compare them with the statements of objectives prepared by faculty members. If an institution's objectives are truly functional it would seem that they should correspond quite closely to those prepared by members of the teaching staff. Certainly if there are wide discrepancies between the two lists there would be serious cause for concern about the validity of the institution's objectives.

At Syracuse University, as a part of its comprehensive self-study, ratings of objectives in general education were made by faculty and students as well as by alumni (53 , pp. 4-13). In one phase of the study faculty members were asked to indicate their objectives in specific courses. In another phase, faculty members ranked the importance of various objectives of general education and indicated how much responsibility their area of instruction assumed for helping students attain each of the objectives. One of the interesting findings of the study was the amount of agreement among students, faculty, and alumni about the relative importance of eighteen objectives of general education.

While not specifically designed to provide a comparison between faculty statements of objectives and the college's published objectives, the Minnesota study reported on by Eckert (18 , p. 14) could have taken this direction. The general procedures would appear to be applicable to the evaluation of objectives, using the technique of comparing faculty conceptions of the institution's objectives with statements of objectives appearing in college publications.

A basic assumption in this approach to evaluation is that the institution's objectives have been stated so clearly that they can be evaluated and that the members of the faculty have thought enough about the objectives of their particular institution to make possible a clear statement of their ideas on the subject. In actual fact, many faculty members, particularly those who have only recently completed their graduate studies, are concerned primarily with their own academic subjects and have not spent a great deal of time thinking about objectives. Very few graduate students spend time in studying philosophies or points of view in higher education. Lacking this background, such faculty members would be handicapped in preparing statements of objectives. In brief, this technique has potentiality but it is not foolproof.

Another possible approach to the comparison of faculty and administration conceptions of school objectives, and one which overcomes the difficulties of having faculty members verbalize institutional objectives, is that recently developed and used by L. R Tucker to study schools of thought about course objectives.

In this study (as yet unpublished), test items for achievement tests in several subject-matter areas were written and submitted to a panel of subject-matter experts, who were asked to rate the items on appropriateness for inclusion in a test of subject-matter achievement. An inverse factor-analysis yielded groups of teachers who selected items of a given type as most appropriate, and thus it was possible to determine the degree to which the various expert critics subscribed to the different objectives represented by the items.

A similar study, making use of school objectives and including both administrators and faculty members as judges, would provide specific measures of the degree of agreement between the faculty and the administration on the various possible objectives.

Evaluation Needs

Refinement of the judgment process. Preceding sections of this chapter have emphasized that regardless of the technique employed in measuring the quality of an institution's objectives, the factor of judgment is involved in every case. It is fairly obvious, therefore, that improvement in the evaluation of educational objectives is dependent in large part on improvement in the effective application of judgment. How can this be accomplished?

One possible line of investigation might stem from the hypothesis that the most important factor in the judgment process is the competence of the judge. Experience, training, and perhaps some difficult-to-define factor which may be termed "judgmental aptitude" appear to determine the competence of a particular judge. The key to the improvement of the judgment process would seem to lie in the careful selection of judges and in training them to perform their particular function.

In an educational situation it would seem that the members of the faculty as a group should be thoroughly briefed on the importance of thinking about objectives. If the subject of objectives has been considered in faculty groups at the departmental level as well as in the faculty as a whole, one could expect to find in each faculty member a background of thinking and experience which would lead to clearer thinking about the functions and purposes of the institution. Since the final list of objectives and the curriculum supporting it must ultimately be approved by the faculty, it would seem that the first step should be that of "educating" (but not indoctrinating) those whose judgment ultimately will determine the issue.

A faculty as a whole is usually too unwieldy a group to consider intensively a topic such as evaluating educational objectives. A smaller group must be selected to carry the main burden of discussion and evaluation; the selection of this group or committee will determine to a large extent the success of the whole venture. The administrator or committee making this selection should be guided, it would seem, by the qualifications of the judges rather than by personalities, representation of strategic areas, or departmental politics. And what these qualifications are would seem to be a fruitful area for investigation.

Once a carefully selected group has been assembled to function as judges, it is essential that there be a thorough briefing on the task to be performed. The nature of the problem, the techniques to be used, the experiences of other institutions, and the ultimate use of the results must be thoroughly covered. Lack of adequate information or understanding of the problem would appear to be one of the most common reasons for inadequate functioning or failure of the judgment process. The importance of properly training the judges can hardly be overemphasized. What constitutes adequate training for group problem-solving would seem to be another significant topic for research.

Once the judges have been selected and prepared for their task, it becomes necessary to choose the techniques which will be employed in gathering and processing the raw data to be used in the evaluation of objectives. Will these raw data be the ideas of the evaluating group, statements of objectives of other institutions, the opinions of faculty, students, and lay leaders, the observation and analysis of the activities of graduates? Will check lists, classification categories, and rating scales be used in the treatment of the data and in arriving at decisions? The answer would seem to be that the best available techniques which are appropriate for the occasion should be used. The behavioral sciences have in recent years developed a good many techniques for sharpening the judgment process. What is needed at the present time is a thorough study of the techniques applicable to the college situation, in which faculty committees function in the role of objectives evaluators. To date we have descriptive accounts of the functioning of college committees but very little information about what determines success in using college committees in the evaluation of objectives.

Improved methods for determining the institution's role. The introductory section of this chapter presents a quotation from Haggerty in which he stresses that each institution should carve out a role for itself and then adhere to it in practice. This is good advice, but how does one determine an institution's role? In the section on "Dimensions of Evaluation" it was stated that an institution's objectives should be appropriate for the clientele and community served. In the next section, "Techniques of Evaluation," it was suggested that study and discussion, the analysis of opinions, follow-up studies of alumni and former students, and comparisons of faculty and published lists of objectives would seem to be the obvious means for carrying out the evaluation process. Of these, the first three seem to possess the most promise for helping in the determination of the institution's role. Any improvements which can be made in these techniques would, therefore, contribute to the improvement in the definition of an institution's function.

One statement can be made with considerable assurance: No one technique can solve the problem of determining an institution's role; indeed, one can hardly think of a problem more complex. Facts and figures can help, but the ultimate questions are in the realm of educational philosophy. Should a junior college teach only terminal courses, or should it offer essentially a college preparatory course? Should a teachers college become an all-purpose college? Should a particular university stress primarily the pure and applied natural sciences, or should it do just as much in the social sciences and the humanities? Should it add a medical college, or should it not? These are among the most important and vital questions with which faculty members, administrators, and governing boards can be confronted, and the arguments are never entirely in favor of one alternative. No one technique or group of techniques can provide the answers. One can say, however, that in finding the answers, the best available facts and figures and the most useful research and statistical tools should be used to provide the raw data for making the final decisions. But, regardless of all that has gone before, the final step in the process must be that of exercising objective, unbiased, and imaginative judgment!

One kind of information which would be very helpful in the proper exercise of judgment would be a compilation of educational objectives from representative samplings of the main types of institutions of higher education. At present these statements of objectives are scattered through many publications of individual colleges and universities. A summary of the type here suggested was prepared several decades ago by Koos and Crawford (38 , pp. 489-509). An up-to-date compilation of objectives of many different institutions, classified by type of institution, would be very useful and might also encourage the evaluation of objectives on the part of those who may not have done much thinking about this subject.

Educational literature contains very few comprehensive and detailed reports of the techniques and procedures used by educational institutions in defining their roles. The problems of rising enrollments in many states are prompting state authorities and governing boards to make comprehensive surveys of the states' needs for higher education and of the part which each type of institution should play in a developing program. Perhaps these studies will lead to published reports which will describe, more fully than has ever been the case in the past, how colleges and universities have defined their roles. Certainly such reports could provide the knowledge which we must have in order to evaluate the extent to which an institution's objectives define its proper role.

Knowledge of what kinds of discussions tend to lead to adequate statements of objectives. It would be a fairly safe guess that the educational objectives of many colleges and universities were formulated without much consultation with all persons and groups directly concerned. The same could be said for the departmental objectives of most colleges and universities. Thus evaluation of the discussion process leading to the formulation of educational objectives would seem to have at least two aspects: first, is there a discussion? and second, is the discussion conducted in an effective manner?

Educators, business executives, military leaders, and specialists in such behavioral sciences as psychology, speech, and sociology have in recent years become very much interested in the study and improvement of the discussion process. So far, not many clear-cut statements of principle can be published. It can be hypothesized, however, that the effectiveness of discussion depends very largely upon the quality of the leadership, the background and experience of the participants, and the atmosphere which prevails in the organization.

The quality of leadership and the background and experience of the participants may be determined largely by the carefulness of the selection process and the adequacy of the educational program arranged for the participants. The matter of atmosphere is more elusive. If a campus has been administered in a highly dictatorial manner, one could hardly expect the faculty to enter effectively into the discussion process. If there is no standing committee structure and no effective system of communication between the administration and the faculty, the basis for carrying on fruitful discussion would seem to be lacking. Perhaps even more important than the administrative and communications structure is the atmosphere, or the lack thereof, of mutual understanding and respect. One hypothesis which might be offered is that an institution which has superior leadership, a well-informed faculty, and a democratic climate of work can be expected to have good discussions of objectives; an institution lacking these qualities will gain little -- even though it goes through the motions of discussion. Much research remains to be done on the relationship between the "atmosphere" of a college or university and the quality of discussions leading to statements of objectives. If measures of this "atmosphere" could be obtained and correlated with ratings of the quality of discussions of objectives, perhaps a beginning could be made toward improving the deliberations which precede the formulation or reformulation of institutional objectives. In addition, of course, one would hope that the findings of the behavioral sciences on techniques for the improvement of discussion would be carefully and continuously studied for applications to the improvement of discussions of educational objectives.

Better correlation between student and institutional objectives. Since the educational program is designed to serve the needs of students, it would seem logical that there should be a correlation between the stated objectives of the institution and the objectives which its students are seeking to achieve. It is not uncommon for colleges to ask students their reasons for coming to college. The emphasis in this sort of questioning is likely to be on vocational objectives. The number of colleges which expect the student to be fairly elaborate in writing out his reasons for wanting to come to college is rather small. The number who ask this question with a view to comparing these responses with the college's list of objectives is even smaller. That most institutions do not systematically compare student and institutional objectives seems regrettable.

The task of correlating student and institutional objectives is, of course, fairly complex. The immaturity of the student, the failure of many students to think about their own objectives, and their lack of adequate background would be serious limitations in the correlation process. It can be argued, for example, that many students don't really know what is best for them. They may think that vocational preparation is the only thing that matters, when preparation for the responsibilities of citizenship may be even more important. Furthermore, lack of correlation between student and institutional objectives might simply reflect inadequate pre-college counseling rather than inappropriateness of the college's objectives. Perhaps this section should be concluded by saying that there is a need for colleges to compare student and institutional objectives, but the results should be interpreted with caution.

Summary of Evaluation Needs

A study of the literature of higher education makes it clear that much has been said and written about educational objectives but that there are few studies which attempt a careful, objective evaluation of them. Few institutions have published studies of their objectives which indicate that the dimensions of evaluation suggested in this report have been considered; very few have applied the best available techniques to the study of their objectives. The most urgent needs in evaluation of objectives are presented in the summary table at the end of this chapter.

It would seem that the greatest need at this time is for institutions to realize that the study of objectives is important. Each institution should ask itself periodically: "What are we doing and what ought we to be doing?" The temptation for institutions as well as individuals is to assume that things are going well and that the objectives of a decade, a quarter century, a half century, or even a century or two ago are still appropriate. The changing times demand reconsideration of objectives. Any means which can be found to stimulate each and every institution to study and revise its objectives would, therefore, be fulfilling a most urgent need.

It would be difficult and perhaps undesirable to suggest that certain specific techniques and procedures should be used in the evaluation and redefinition of objectives. The critical requirements technique of Flanagan, the "diary study" of Stephens College, and the questionnaire and interview studies of the University of Minnesota are interesting and promising methods for use in the study of objectives and should be repeated from time to time. Some studies of this kind would be of interest to all types of colleges; others could be planned which would be of value to one type of school, e.g., engineering, law, or medicine. But most important of all is the problem of stimulating all institutions to think and talk about their objectives, particularly in this time of rapid change in instructional responsibilities and of skyrocketing enrollments. A program directed toward the study of objectives on the part of all institutions would indeed be timely and worth while.

Summary Table for Chapter II

Evaluation of Institutional Objectives

Evaluation Problems →	Appropriateness of the Objectives	Precision of the Statement of Objectives	Adequacy of the Process of Formulating Objectives	Extent to Which the Objectives Have Been Elaborated	Faculty and Administration Concern about Objectives
Available or Existing Measures and Techniques for Evaluation	1. Group study and discussion 2. Individual opinions and ratings 3. Follow up* 4. Comparison of faculty conception of the objectives with published statements	1. Group study and discussion 2. Individual ratings and opinions	1. Ratings and opinions	1. Group study and discussion 2. Opinions and ratings	1. Opinions and ratings
Measurement Needs	1. Improved techniques for recording and combining judgments 2. Method for obtaining information on needs of community and prospective students of the college 3. Instrument for obtaining faculty conceptions of objectives	1. Improved techniques for recording and combining judgments 2. Ratings or objective measures of the usefulness of the stated objectives to: a. Faculty for guiding instruction b. High school seniors and their parents for selecting a college	1. Check list of procedures	1. Ratings or objective measures of the usefulness of the elaborated objectives to: a. Faculty for guiding instruction b. Students for selecting a major	1. Scale for measuring attitude toward objectives 2. Indices of amount of faculty and administrative time devoted to consideration of objectives
Available Interpretive Data					
Needed Interpretive Data	1. Normative data for the ratings 2. Appropriate background information describing the community and prospective students	1. Normative data for ratings 2. Normative data for objective measures	1. Data on how colleges formulate objectives 2. Information on what constitutes a good method of formulating objectives	1. Normative data on all measures	1. Normative data

---

*See Summary Table for Chapter VIII.

PRECEDING PAGE BLANK- NOT FILMED

Chapter III

THE CURRICULUM

Introductory Statement

The curriculum is the heart of any institution of higher learning. Modern colleges and universities perform many valuable services in our society, but in the final analysis they must be judged primarily by what they teach. Many of the activities carried on by institutions of higher learning and many of the services performed by them could, if necessary, be shifted to other agencies, but the activities of the classroom and laboratory could hardly be duplicated elsewhere. It is because teaching is central that college faculties and administrative officers worry about the curriculum. Any report on the evaluation of colleges and universities should, therefore, concern itself with methods for the evaluation of the curriculum.

Evaluation Problems

The purpose of this section is to describe or define the dimensions along which various college curricula appear to differ and which will enable one to place these curricula on a scale ranging from high to low quality. This discussion will deal exclusively with the dimensions of measurement; the techniques for making the measurements will be described in the next section.

Implementation of the institution's objectives. The previous chapter emphasized the fundamental role of objectives in determining the program and activities of a college or university. It would seem, therefore, that first and foremost the quality of a curriculum should be judged by determining how well it implements the institution's objectives. Chapter II stated that college catalogs typically present eloquent statements of the objectives which the institution is seeking to achieve, and that the college's program of activities does not always appear to be directed to the realization of the stated objectives. The curriculum is, perhaps, the best and most crucial test of the extent to which institutional objectives are followed. If a college's objectives are clearly stated, and if its curriculum is definitely and discernibly related to these objectives, it would seem that such a college should be placed on the plus side of the college evaluation scale; conversely, a college whose curriculum is unrelated to its stated objectives should be placed on the minus side of the evaluation scale.

Recognition of the instructional needs of students. A second problem in the evaluation of the curriculum is the adequacy with which it recognizes the instructional needs of its student body. A private liberal arts college and a multi-purpose state university, for example, may have quite different student populations. Some institutions of the first type are highly selective in their admissions policies and emphasize general or liberal education as opposed to professional training. Many state universities, on the other hand, admit almost without restriction graduates from their own high schools and generally provide pre-professional and professional education of various kinds. The curricula of these two types of institutions should reflect, in level of courses taught and in range of offerings, awareness that their student populations are different. There is perhaps, a natural tendency for institutions to promise too much, thereby disappointing a certain number of their students. On the other hand, the range of courses offered may be considerably greater than the instructional needs of students. Both extremes are undesirable; the highest rating should go to the institution which has accurately assessed the instructional needs of its students (perhaps by making a careful inventory of their educational interests, vocational plans, and citizenship responsibilities) and has tailored its curriculum to meet those needs.

Recognition of individual differences among students. A third problem in the evaluation of the curriculum is the extent to which it recognizes individual differences in academic aptitude and previous preparation. The subject of individual differences has been dramatized during the past forty years by many studies in psychology, education, and sociology. Despite the many studies, despite the lip service given to the importance of recognizing individual differences, many college courses are taught without serious consideration being given to this problem. Even in such required courses as freshman English, mathematics, and "core" courses, it is not uncommon for individual differences in student aptitudes and previous preparation to be largely ignored. Fortunately, there are institutions which have given a good deal of attention to this problem by planning extensive testing programs for entering students and offering instruction based on various levels of aptitude and preparation. Naturally this problem is related to the admissions policy of the institution, but even in the most carefully selected student body there will be important aptitude and background differences. The extent to which the curriculum recognizes these differences is one means of determining its adequacy or quality.

Articulation with curricula of other educational units. A fourth evaluation problem is the extent to which the curriculum recognizes the education which has gone before and the education which will follow in another school, or in life. The American college is a unit in a total program of formal education which may cover a period of twenty or more years for some professional workers. There are problems of articulation with the secondary school on the one hand and with the graduate or professional school on the other. Overlap of curricula at either end will result in loss of time and frequently in loss of interest on the part of students. The problem is greatly complicated by the fact that the college draws its clientele from many different secondary schools and its graduates enter many types of professional schools and graduate departments of instruction. Difficult as the problem is, the curriculum which is most effectively articulated with both the secondary school and the graduate or professional school is the most successful.

Meeting the needs of society. A fifth type of evaluation problem is the extent to which the curriculum meets the needs of society. This criterion is probably the most important and the most difficult to measure. The literature of education is replete with books and articles setting forth the thesis that the school system must meet the needs of society, that every effort must be made to determine those needs and to tailor a program of instruction for the young accordingly. Today much is being said about the shortcomings of the elementary school in teaching the basic skills of communication and calculation, and of the failure of the secondary school to provide students with the facts and skills which they need to carry on college studies; colleges and universities are frequently criticized for training specialists lacking the liberal education which should prepare them for the general responsibilities of citizenship in a democracy. Whether a curriculum really meets the needs of a society is a very important question. Difficult as it may be to determine what are the needs of society, it is inescapable that the quality of a curriculum will be judged in considerable part by how well it appears to meet society's needs.

Techniques of Evaluation

In the preceding section five dimensions of curricular evaluation have been discussed. Quite possibly there are others, but these five would seem to be among the most significant. How does one determine the adequacy or quality of a particular curriculum with respect to these dimensions? Some of the many approaches to curricular evaluation have dealt with the curricula of entire schools, others have dealt with special facets of the total curriculum, and still others have attempted to evaluate the effect of education as a whole, no attempt being made to differentiate between the curricular and the extra-curricular. This section concerns itself with a description and discussion of some of the most commonly used techniques for curricular evaluation.

Evaluation by opinions. Perhaps the oldest and most widely used technique for curriculum evaluation is that of soliciting opinions from faculty members, students, parents, graduate and professional school faculty members and administrators, employers of college students, etc. In some cases, of course, these opinions have been offered without any very formal means of gathering or analyzing them; in others, much time has been spent in devising questionnaires or rating scales and in very complete analyses of the results. A survey of the literature soon reveals that many opinions have been expressed about the curriculum; indeed, it would seem that no study of a curriculum would be complete without some use being made of opinions as a means of evaluation.

An illustration of this approach to curricular evaluation and development is described by Heaton and Koopman (27) in their account of a study conducted at Central State Teachers College, Mount Pleasant, Michigan. The faculty made an extensive study of the functional needs of its students by surveying student opinions and by sending questionnaires to graduates of the college, asking them to record the abilities, knowledges, understandings, and appreciations needed in their various adult relationships. Interviews were also held with parents of students and with school administrators. The faculty then analyzed their findings and made the necessary curricular revisions and administrative reorganization.

Another study reporting the results of a systematic survey of student opinion regarding an entire curriculum is that of Eckert (17 , pp. 160-190). Undergraduates and former students enrolled in the general college of the University of Minnesota were asked by questionnaires and interviews to express their views on the individual courses and on the college program as a whole. Of particular interest is a question which asked whether students would recommend the general education program to a younger sister or brother. Answers to this question as well as to others pertaining to individual courses seem to indicate that students can provide useful evaluations of a curriculum. Students, like most other groups of raters, are likely to be overgenerous in their comments, but a great deal of useful information can be obtained if they are given an opportunity to comment rather fully on their evaluations. Student opinions are, of course, widely used in the evaluation of quality of instruction, examinations, visual aids, textbooks, and other aspects of the actual teaching process.

Evaluation by opinions, or ratings, has been extensively used by accrediting teams. The North Central Association (49 , pp. 178-179), for example, constructed a rating card for measuring the important aspects of a college curriculum. Weights were given for the presence or absence, or degress of quality, of the major aspects of the curriculum. Basically this is the technique of utilizing expert opinions, the experts being a team of outside raters. The results of the North Central study show that these ratings by experts are positively correlated with the over-all quality of the school. If such ratings could be collected on a large-scale basis and reported in tables of norms, a significant contribution would be made to curriculum evaluation. The merits and demerits of this approach will be considered more fully in the next division of this chapter.

Study and discussion. Closely related to the opinion questionnaire is the method of discussion, usually by committees of various kinds. Practically every college has a curriculum committee and from time to time appoints special committees to study various aspects of the curriculum and to make recommendations concerning possible changes. Much of the total meeting time of college faculties is devoted to discussion of curriculum, particularly of courses required of all students. Indeed, in the final analysis, the curriculum adopted by a college and the changes which may be made from time to time are the direct result of faculty discussion.

The use of the discussion method in evaluating the curriculum is well illustrated in the development of the new curriculum at the State University of Iowa in 1944 (47). Committees were appointed to deal with particular aspects of the curriculum -- social studies, history, science, foreign languages, fine arts, literature, skills, and abilities. Each of these committees had many meetings over a period of two years and reported to a sub-committee of the standing committee on curriculum entitled "Steering Committee on Curriculum." Each of these committees had at least seven or eight members; hence a large part of the faculty served on one or more. The general procedure in each committee was to make a thorough analysis of the area of the curriculum assigned to it (e.g., science), to discuss what ought to be done, and finally to make recommendations to the Steering Committee on Curriculum. The latter committee prepared a summary report and a set of recommendations which were submitted to the faculty and resulted in the adoption of a new curriculum in 1944.

The procedures used at the State University of Iowa have been used at many other institutions both before and since 1944. It is what one might call the common-sense approach to curriculum evaluation and revision. The method is effective in proportion to the interest and energy of the committee members, especially the chairman, and to the clearness with which the function of each committee has been defined.

Follow-up studies. The quality of industrial products is usually judged by how well they perform in the hands of consumers. Automobile manufactures, for example, stress the satisfaction of their customers with their particular make of automobile. Since this practice of judging quality by product performance is so widely used in other areas, it has been suggested by many that the quality of the educational process should also be judged by its product, the performance of its graduates or former students. A great many studies have been made of the "products" of various courses, curricula, particular institutions, and colleges in general. All these studies have in common the idea that the performance or behavior of those who have been exposed to a particular type of education can be used to judge the effectiveness of that educational process. These studies are generally referred to as "follow-up studies."

One of the pioneer, influential follow-up studies was undertaken by Pace at the University of Minnesota (52). Pace studied 951 former Minnesota students, asking each to fill out an extensive questionnaire which was concerned somewhat with opinions but primarily with the former students' current activities, personal, vocational, and in the home and community. From the questionnaire data it was possible to compare graduates and non-graduates with respect to their current interests, opinions, reading habits, community activities, and so forth. While this study was limited to one institution, the methods employed have general applicability to the evaluation of curriculum. Indeed, Pace's results were used in the refinement of the work of the General College of the University of Minnesota and in connection with the general education program of the College of Science, Literature and Arts.

One of the most recent and most widely publicized follow-up studies is "They Went to College" sponsored by Time, Inc. (26). Several sections of this report deal with matters which pertain to the curriculum. One chapter, for example, is entitled "The A.B. and the Specialist"; another is "Princeton versus Podunk"; still another, "What Students Think About College." This study represents only a very general approach to the study of the curriculum. It is worth mention primarily because of the care with which the sample was drawn and the comprehensiveness of the study. Any future investigation of this scope would profit from the experience gained in this particular project.

The chief defects of the study, particularly with reference to predictor variables and criterion measures, are discussed in Chapter VIII.

One of the best-known follow-up studies in higher education is based upon the Eight-Year Study sponsored by the Commission on the Relation of School to College and entitled Did They Succeed in College? (10). This study was concerned with the success in college of graduates of thirty secondary schools who freed their students from the usual requirements and permitted them to follow, under guidance, their own study inclinations. A considerable number of colleges agreed to accept, upon the recommendation of the secondary school, these "non-conforming" students, despite the fact that they did not meet the usual entrance requirements. A total of 1475 of these special students was admitted and matched with a control group who had followed the traditional preparatory-school curriculum. One group of students was followed over four years of college, one over three, one over two, and the last group was studied for just one year. The criteria used in evaluating these college students were as follows: intellectual competence, cultural development, practical competence, philosophy of life, character traits, emotional balance, social fitness, sensitivity to social problems, and physical fitness. The sources of evidence concerning each of these criteria were tests, interviews, questionnaires, records of various kinds, reports of instructors, placement heads, employers, supervisors of activities, and parents. The tests used were novel in scope, including, for example, a scale of beliefs, tests on social usage or ability, contemporary affairs tests, personality tests, and subject-matter placement tests. Questionnaire and interview data were studied by trained raters or reviewers and the mass of reported or observational data was reduced to a set of ratings, usually on a scale of 1 to 5. The ultimate objective in nearly all phases of the study was to make quantitative comparisons of the "progressive" and "traditional" students, the data being gathered from a considerable number of colleges and universities.

While the study did not reveal striking differences between the "progressive" and "traditional" students, it did make a considerable contribution to the literature of curriculum evaluation. The experimental populations were carefully selected, many new measuring instruments were developed, and emphasis was placed on criteria of achievement other than grades in courses. More complete data were gathered on these two groups of college students than had been gathered on any groups of comparable size in the whole previous history of higher education. Last but not least, the study demonstrated that secondary schools and colleges can cooperate in large-scale curricular studies. From a practical point of view the study, perhaps, stimulated colleges to relax formal entrance requirements and to attach greater weight to student ability and promise as the major factors in determining qualification for college admission. But most of all the study blazed a new trail in the whole field of curriculum evaluation. For this it is primarily remembered.

Comparison with other institutions. While there is a great deal of variety in American colleges and universities, there is also considerable similarity, particularly among institutions which purport to perform the same kinds of educational tasks, e.g., liberal arts colleges or engineering colleges. For this reason a good deal can be gained from comparative studies made on a systematic basis. College catalogs describe the courses of study and curricular requirements of particular institutions, but frequently these statements are not strictly comparable, thus making comparisons difficult. In order to make helpful comparisons it is usually necessary to lift the data from college catalogs, supplement it with questionnaire data, and then present it in tabular form.

A study illustrating the comparative approach to curriculum investigation was made by the College of Liberal Arts, State University of Iowa, in preparation for the 1955 meeting of the Deans of Liberal Arts in State Universities of the Mississippi Valley (84). Each of these institutions prepared a statement of its graduation requirements for the Bachelor of Arts degree, using the following outline: (1) Basic skills, including requirements in reading, writing, speaking, and calculating; (2) Area requirements in humanities, natural science, and social science; (3) Foreign language requirements; (4) Major, or area of concentration, requirements; (5) Military science requirements; (6) Electives; (7) Qualitative requirements for the degree; (8) Residence requirement. The compilation of these requirements enables any institution to compare its own requirements for the bachelor of arts degree with those of sister institutions. Comparison of a particular school's practices with those of other institutions cannot by itself lead to the solution of curricular problems; it can, however, provide very useful information for discussion and for the making of recommendations to a college faculty.

Educational and psychological tests. A great many studies have presented the results of educational and psychological tests as measures of the success or lack of success of various types of courses and curricula. Notable among these studies are those of Chamberlin and others (10), Dressel and Mayhew (14), and Eckert (17). In the study by Chamberlin and others (10), achievement tests, a scale of beliefs, and a test of social usage were used to measure the achievement or development of matched groups of college students, one group having followed a "progressive" secondary school curriculum and the other a "traditional" program. Pre-tests and post-tests were given in a number of instances in order to measure growth over a period of time and to make possible comparisons between the matched groups.

Dressel and Mayhew's (14) study dealt with the measurement of general education outcomes, including the areas of social science, communications, science, humanities, critical thinking, and attitudes.

The principal technique was that of the pre- and post-testing of students enrolled in the schools participating in the study. While some results are presented for groups of students having different types of educational experience, the method of matched groups was not generally employed. Committees of faculty members and test experts developed the testing instruments. While the study as a whole does not reveal very striking results in measuring general education outcomes, it does seem to be one of the best of the cooperative studies in the field of curricular evaluation. Some of the problems encountered in conducting this type of study, particularly those of measuring growth and of correcting for differences in initial student ability or achievement, are discussed in Chapter V.

Eckert (18), in her study, also used achievement, personality, and special tests of various kinds to measure outcomes of instruction in the General College of the University of Minnesota. In some cases comparisons were made with the performance of students in other programs such as the College of Science, Literature and the Arts as well as with college students in general. Among the testing instruments used were the Cooperative English Test, the Minnesota Reading Test, the Personal Living Test, the Minnesota Personality Scale, the Bell Adjustment Inventory, the Home and Family Relations Test, a vocational orientation test, the Socio-Civic Affairs Test, the Cooperative Contemporary Affairs Test, and the Pace Situations Test. Other measures of the effectiveness of the general education program were also used, such as interviews, questionnaires, check lists, and ratings, but the testing program constituted a most important part of the total evaluation project. Perhaps few colleges in the country have so thoroughly evaluated their curricula. To be sure, many of the obtained data have shortcomings; nevertheless, an evaluation of this scope cannot help but have a significant effect upon all those connected with the institution -- faculty, students, and administration.

Persistence studies. It was mentioned earlier that many businesses and industries use the satisfied customer as one very important measure of the quality of their products. Applying this measurement technique to education, one might argue that the continuation in school of those who enroll as freshmen should be one criterion of satisfaction with the curriculum. In state universities without selective admissions policies about 30 percent of the entering freshman class do not enroll for the sophomore year; an additional 10 to 15 percent fail to enroll for the junior year, and only about 40 to 45 percent finally receive the degree. The reasons for withdrawing from school are very complex, and care must be exercised lest causes for withdrawal which actually reside in the student himself be attributed to the institution. Nevertheless, if a particular college has a drop-out rate significantly greater than that of others of its type, it would seem logical to ask whether the curriculum of the school is a factor. There are no studies in the literature which show a direct relationship between quality of curriculum and student persistence, but this is largely because it is difficult to single out from among many factors the impact of the curriculum upon withdrawal rate. Logically it would seem that the student's satisfaction with the curriculum would be one important factor determining the length of his stay in a particular school.

Closely related to withdrawal rate is enrollment itself. The North Central Association's study of institutional evaluation (49 , pp. 126-131), suggested that the rate of growth of a college should be considered as one factor indicating its total quality. This, again, would seem logical. If all or most other colleges in a particular group are growing but the enrollment of a particular school remains static or declines (assuming it has staff and facilities for a larger enrollment and assuming that the institution has not deliberately determined to hold its enrollment constant or even to reduce it), it would seem that one might question the quality of its curricular offerings as well as other facets of its total program. The same principle would seem to apply to the number of students applying for admission. While the factors of withdrawal rate, enrollment, and number of applicants cannot be taken by themselves as indices of institutional quality, they can invite questions which may result in the revelation of important facts about the school and its curriculum.

Measurement of duplication. Many critics of the American educational system (79,82) have commented on the duplication among the various divisions (elementary, secondary, college, graduate and professional) of the system and the duplication of course materials within a particular division, for instance, the number of statistics courses offered in a typical university. Some duplication may be necessary and desirable, but we have probably erred on the side of too much duplication rather than too little. The techniques for measuring duplication include those of opinion seeking and study (including observation) and discussion already discussed as techniques of curriculum evaluation. For example, some studies (82 , Appendix, p. 133) have invited students to comment on the degree of duplication between secondary school and college work or between courses in the college curriculum. The results of such studies, while they must be interpreted with caution, are helpful to college faculties and curriculum experts. The student may not be a very good judge as to just how much duplication of subject-matter is wise, but he can report on the presence or absence of it and the extent to which he feels that it has affected his educational progress.

Individuals and faculty committees can, of course, make direct inspections of course titles or of the syllabi of various courses and arrive at estimates of the amount of duplication between different divisions of the educational system and between courses within the same system. The writers of General Education in School and College report, for example, that they found a striking similarity between the twelfth-grade English course in three preparatory schools and the standard freshman English course offered at two of the universities participating in their study. In the sciences it was found that almost half of the students included in their "transition study," took in college the beginner's course in the same science they had taken in high school, and many repeated two of their sciences when they enrolled in college.

Another kind of evidence which would throw light on the articulation between high school and college is found in studies of the subsequent achievement of students admitted to college without high school graduation or with advanced standing. One hypothesis is that if such students experience little difficulty in their college work, either duplication exists or some work of the secondary school is not essential for the college courses taken by students. A recent summary of the effects of acceleration has been prepared by Flesher and Pressey (21), and a report on granting advanced standing by examination has been published by Jones and Ortner (35).

Evaluation by Experimentation. Despite all that has been said and written about the problems of curriculum construction and the merits or demerits of particular courses or curricular organizations, there has been little well-designed curriculum experimentation on the college level, particularly of the sort which would be useful in curricular evaluation. The outcomes of education are very difficult to measure, it is difficult to conduct experiments in an ongoing program of education, and many faculty members lack confidence in the experimental approach, and therefore, though a number of studies have been made of the teaching process, studies of different curricular organizations are fairly rare.

One example of an experiment involving the measurement of outcomes of instruction using different types of courses was conducted at the State University of Iowa in a self-study financed in part by the Fund for the Advancement of Education (16). This study was concerned primarily with whether so-called "core courses" are any more successful than departmental introductory courses in achieving the goals of general education. The study was conducted in two of the four divisions of the curriculum, social science and natural science. Entering students were given the option of taking either core or departmental courses, and at the end of the year all were required to take the comprehensive examination which applied in their case -- social science, biological science, or physical science. These examinations were constructed by faculty committees which tried to construct examinations measuring general education (as opposed to specialized knowledge) in their particular areas. Using the technique of covariance, all groups were equated for ability on the basis of their composite scores in the entrance examinations. Primarily, comparisons were made between students who had had the eight-hour core course in the area and those who had had eight hours of regular departmental introductory courses in the area, e.g., four hours of physics plus four hours of chemistry. Most of the differences favored those who had had the core courses, but in general the differences were not statistically significant.

In any experiment such as that described above the nature of the comprehensive examinations is crucial. For this reason the designers of the above study also tested seniors, using the general tests of the Graduate Record Examination, and then analyzed the backgrounds of the students to set up groups who had had and who had not had core courses. In this case also the students who had had core courses, in general, did slightly better than those who had not had such courses (all groups were equated for differences in ability), but the differences were small. In other words, the two parts of the study were in substantial agreement. Partly on the basis of these findings the faculty decided to give students the option of taking either core or "approved" departmental courses; that is, the faculty accepted the findings as constituting an evaluation of a curricular requirement which had been in existence for ten years and made the change which appeared to be indicated.*

The State University of Iowa study of core courses was conducted in a single institution. The evaluation of the results of such a study would be enhanced by the availability of results of similar studies at other institutions. In other words, the value to curriculum evaluation of the results of experimental studies increases as the number of studies is increased, particularly among institutions of similar educational programs and purposes.

Evaluation Needs

The preceding sections have been concerned with evaluation problems, or "dimensions," and the techniques which have been used in evaluation of the college curriculum. The published literature on the curriculum is very extensive. Most of the articles and books on the curriculum are of the "non-experimental" type. This is quite understandable, for in the area of curriculum construction and evaluation, exposition and argumentation have a real place; value judgments are basic in many curriculum decisions. At the same time it would seem that there should be a place, perhaps a larger place, for the objectifying of data and the application of the experimental approach. In order to do this, better measuring techniques and instruments, better experimental designs, and better methods of analyzing data are needed. This section considers these needs.

Better use of opinions or judgment. The opinions of faculty members, administrators, students, employers, alumni, and teams of experts will always play an important role in the evaluation of the college curriculum. This means, in effect, that one of the main needs in the improvement of curriculum evaluation is that of the sharpening or objectifying of subjective judgment. Whenever we use opinions as measures of anything we are relying upon the human mind (coupled with some emotions perhaps) to function as a measuring instrument? How can one improve the quality of that instrument.

---

*See also a study by Bond (8) described in Chapter V. This study illustrates the experimental evaluation of a unit of subject-matter involving both content and method.

In a sense, all that has been written about opinions and ratings in the fields of psychology, education, and sociology is applicable to this problem. The instrumental and statistical problems in rating and in opinion polling are adequately treated in textbooks dealing with these subjects. The curriculum evaluator who would use opinions or ratings in the evaluation process would be well advised to familiarize himself with the extensive literature in this technical field.

The main problems encountered in using opinions and ratings in curriculum evaluation would seem to be: (1) securing judges who have the proper background for making valid and reliable judgments, (2) training and instructing judges to apply their knowledge and experience effectively, (3) recording judgments so they can be studied and analyzed statistically, (4) compiling norms of these judgments so they can be applied in new situations.

The North Central Association study mentioned earlier made a serious attempt to meet these problems. Persons thoroughly familiar with college curricula were asked to do the rating. The rating procedures were clearly explained and the objectives of the total study emphasized. Rating scales or score cards were carefully devised to facilitate the recording of judgments, and, finally, tables of norms based upon these ratings were constructed. It would appear that more studies of this kind are needed to supplement the local studies which have been so important in the improvement of college curricula.

Better experimental evaluations. The number of truly experimental studies of the college curriculum is not large, and the designs of experimental studies reported in the literature often leave much to be desired. While the results of most studies are analyzed by statistical techniques which assume random assignment of cases to each experimental treatment, it is comparatively rare to find studies in which students have in fact been assigned on a random basis to particular curricular treatments. Of course, there are practical reasons for this omission. Most college faculty members and administrators feel that they should do the best possible educational job for students and should not subject students to curricular experiments which may be of doubtful value. Arbitrary assignment to particular curricular treatments conflicts with the "free elective" system of course selection. And even if the decision were made to conduct an experiment and to assign students to different curricular treatments on a random basis, many colleges would find it difficult to find faculty members who could teach the different groups. While the application of modern statistical techniques, e.g., analysis of covariance, can reduce some of the practical difficulties of the experimental approach to curricular evaluation, the experimental method still requires considerable planning and some change in the normal routine. There are, in other words, a number of fairly plausible reasons for the scarcity of ideally designed experimental studies of the curriculum.

Since the experimental design problems are quite formidable, there is an understandable reluctance on the part of many colleges to undertake curricular experimentation. This reluctance, however, may be based more upon rationalization than upon a valid appraisal of the situation. Colleges and universities, like other institutions, are inclined to continue with the tried and true rather than to turn to something new and different.

Paradoxically, the nature of the educational process is so mysterious that colleges could safely hazard more curricular experimentation. Much has been written about the virtues and shortcomings of "interdepartmental, basic concept" general education courses as contrasted with "narrow, specialized departmental courses"; one might think that the student exposed to one type of course would get a "perfect" general education while the student exposed to the other type would be very unlikely to get any general education at all. In such a climate of opinion the chances are slight that students will be given an option as to which course they would prefer. Yet few, if any, studies of the curriculum show that disaster will overtake the student if he follows one curriculum and complete success if he follows the other. Most of these studies leave room for doubt; students required to follow different courses of study aimed at the same objectives would probably be done no great disservice. Reckless or pointless experimentation is naturally undesirable, but a total lack of experimentation is even more to be deplored. There is enough uncertainty about the effectiveness of different curricular treatments of subject-matter, and enough evidence of the adjustability or flexibility of the college student, to indicate that a particular student does not lose all if he is included in a curricular group which seems to learn less than another group which followed a different course of study. Even though the educational results obtained with one or more groups included in an experiment should be below average, the long-term benefits for future generations of students make such experimentation highly desirable. Soundly conceived experimentation, utilizing proper experimental designs, can contribute greatly to the evaluation of the college curriculum. If large numbers of colleges and universities were to experiment with various features of the curriculum, data would become available which could be utilized by self-study committees and outside judges in determining the strengths and weaknesses of particular institutions.

In the evaluation of curricular experiments it is, of course, important to account for all the variables possible, so that the findings will be fair and reliable. For example, attention should be given to the background and ability of the students in the courses compared, to their purposes in taking the course, to the personality of the teacher, and even to the hour of the day in which the course is offered. The instructional techniques employed will likewise have a bearing on the outcomes. Because of these many variables, some unpredictable, the evaluation of curricular experiments is made difficult; nevertheless, it is important that experimentation be carried forward even though the conclusions must be interpreted with caution.

Better information about curricular practices. American colleges and universities show a good deal of variation in their approaches to the higher education of the student. This is true of institutions of one type as well as of the whole group. It has been said by many leading educators that this diversity is healthy and reflects vigor and originality of thinking about education on the college level. A good many arguments could be presented supporting this point of view.

Diversity of educational practice is much to be desired, and it is likely that particular schools could profit by knowing more about the policies and practices of their sister institutions. A study of college catalogs provides useful facts about the practices of other schools but does not provide the comparative data which are really needed. What are the requirements for the Bachelor of Arts degree in state universities, in four-year non-denominational liberal arts colleges, in denominational colleges and state colleges? What are the requirements for the Bachelor of Science degree in these institutions? What are the requirements for specialized degrees of all kinds? Information of this sort, if gathered on a systematic basis and presented in readily comprehensible form, would be extremely helpful and would greatly facilitate the evaluation of the curricular offerings of a particular institution.

Better follow-up studies. The college alumnus or former student is in a unique position to provide very useful information in the evaluation of college programs, including the curriculum. This is well illustrated in the Minnesota and Syracuse studies and the comprehensive follow-up study conducted by Time, Inc.

What is needed now is a refinement of the follow-up study, using opinion polling of former students and carefully assessing their current behavior, as recommended by Pace (50). Previous studies have provided a great deal of useful data about college graduates, how well they succeed vocationally, to what extent they accept their citizenship responsibilities, what they think and hope and feel, what they believe college did for them, and what else could have been done for them while in college. What would happen if college graduates were presented with a transcript of their records five, ten, and twenty-five years after graduation? How would they evaluate their courses? Would their opinions change with time? How would they change their programs of study if they had a chance to "do it over again"? A combination of questionnaires, activity scales, and interviews would seem to be the best research tools -- although it would also be very interesting and helpful to have achievement test data at these times. If in drawing up the sample a very wide variety of occupational groups could be included, the study would have important values not only for general education but also for specialized education.

Included in a follow-up study should be a representative sampling of those who entered the institution as freshmen but dropped out of school before graduating. In the case of those going to other institutions it would be highly desirable to get an evaluation of the work done at each institution and also to get a measure of the satisfaction with the change in institutions. Very few institutions have made complete studies of former students as well as graduates. It would indeed be highly informative as well as interesting to do so.

While it would involve formidable obstacles, it would seem very valuable to include in the experimental population a representative sampling of high school graduates not going on to college but matching in ability those who did attend college. If enough information could be given about particular courses, it would be most interesting to see how these men and women would evaluate the possible usefulness of these courses. As stated above, the obstacles to be overcome would be very great indeed, but such a study would seem to be well worth a try.

From the standpoint of the curriculum evaluator, the follow-up study has interesting and challenging possibilities -- but also limitations. Some of these were mentioned in the previous chapter and others will be mentioned in later chapters, particularly Chapter VIII. One common shortcoming of follow-up studies is that too little attention is given to the problem of how to measure success in a vocation, as a citizen, as a person, and as a member of a family. In the planning of future follow-up studies the definition of what is meant by "success" should receive very careful consideration.

More and better studies of general vs. specialized courses. One of the favorite criticisms of college curricula is that too many courses are offered, increasing the cost without increasing proportionately the quality of education. Indeed, the critics of the specialized course often argue that the multiplicity of courses results in an inferior education. Some faculty members, on the other hand, would argue that the purposes of education are better served by teaching specialized subject-matter. They would argue particularly that this type of course is needed by those students going on to graduate and professional schools.

A thorough experimental study measuring the relative effectiveness of these two approaches would be of great interest to educators everywhere. What would happen if the number of courses offered on the undergraduate level in a particular department were reduced by 10, 15, or even 25 percent? Would students taking the specialized courses do any better in graduate school than those taking a core of fundamental courses only? How well would these courses serve the purposes of general education?

Planning a study to determine the relative effectiveness of these two approaches would be difficult. Ideally, one should have two departments in each subject -- e.g., English, political science, psychology -- with half the students going through one department and the other half going through the other department. Instruments would have to be constructed to measure what a competent group of authorities felt the student should know about his subject at the end of his undergraduate course. A similar test-construction job would have to be done to measure the outcomes of general education. While not absolutely essential, it would also be desirable to determine the success of these students in graduate and professional schools. Since most institutions are concerned about the costs of education and the relative merits of general and specialized approaches to education, it would seem that this study could serve a very important function.

Summary of Evaluation Needs

The status of curriculum evaluation on the college level is both encouraging and discouraging. A number of significant studies have been reported during the past twenty years, and important contributions have been made to the scientific appraisal of the curriculum; the Eight-Year Study, the work of the General College of the University of Minnesota, the general education study by Dressel and Mayhew, and the study sponsored by the American Council on Education are most noteworthy.

The summary statement written by Eckert (14, pp. 250-278) concerning curricular research in general education applies equally well to the whole field of higher education. Good, solid research is conspicious by its absence. In a sense this is not surprising, for research in this field is difficult to design and carry out. It should also be said that the planning and evaluation of a curriculum must always be, to a considerable extent, subjective. Nevertheless it seems almost self-evident that studies can and should be undertaken which will add to our knowledge about the merits and demerits, the strengths and weaknesses of various types of colleges. The main areas of need are summarized in the accompanying chart.

One of the major problems encountered in curricular evaluation is the diversity of curricula found in American colleges and universities. Strong arguments have appeared in the literature pointing out the strengths which are realized in this type of situation. It is also fairly obvious that any attempt to standardize curricular practice would meet with the strongest opposition. At this time and perhaps at any future time, therefore, it would seem inappropriate to attempt to bring about greater uniformity in college curricula.

Perhaps the greatest need in the area of curricular evaluation is that of the development of testing instruments which will measure with high validity the general outcomes of instruction If it were agreed that all college students should achieve a certain minimum level of performance in the communication skills and mathematics, should have some knowledge and appreciation of each of the major fields of knowledge -- literature, fine arts, history, philosophy and religion, social science, biological science, and physical science -- should have achieved proficiency in at least one foreign language, and should have above average knowledge in some one field, it would be very helpful if instruments to measure these outcomes, plus the normative data to interpret the results, were available. In addition, it would be very helpful to have tests which would measure achievement following one year of study in a field, two years, three years, and four years. While it would increase the magnitude of the task, it would be desirable to have tests for each major department of knowledge as well as for the general areas mentioned above -- and, in addition, to have tests measuring the general outcomes of instruction such as the scientific attitude, respect for truth, and the like.

It is, of course, well known that a great many tests resembling those just described are already available. Notable among the tests now ready for use are the Cooperative tests and the Graduate Record Examination. Unfortunately, many college faculties are unwilling to accept available tests and to use them to measure the outcomes of instruction in their courses. This situation will probably continue as long as the present diversity of curricular offerings continues.

Despite the difficulties which would be involved, it would seem desirable to plan to do the following:

(1) Construct subject-matter tests of a validity which would satisfy representative faculty groups.

(2) Administer these tests to students (a) with varying amounts of work in the field or department and (b) enrolled in various types of colleges.

(3) Make available these tests and norms for use by colleges in carrying on curricular studies in their own situations. At present, test norms for various types of colleges are available in only a few fields. Individual colleges and universities could do a better job of curriculum evaluation if test norms were more clearly defined, based upon the performance of students in representative colleges of the class to which the particular college feels it belongs, and published for different levels of student ability, possibly by presenting norms in the form of regression lines of achievement on ability.

In brief, refinement of techniques and methods of investigation are most urgently needed for curricular evaluation. We have some general ideas about the goodness or badness of various curricular practices but nothing very precise.

Furthermore, we don't really have very clear notions as to just how much the typical freshman, sophomore, junior, senior or graduate knows about a particular field. Neither do we know whether one type of curriculum (e.g., departmental or interdepartmental courses) is necessarily better than another. We know that colleges differ and that students differ, but the detailed information as to why and how much they differ is lacking. The next great forward steps would seem, therefore, to be the refinement of measurement techniques and the assessment of outcomes of instruction, which to date have largely escaped our measuring instruments.

PRECEDING PAGE BLANK– NOT FILMED

Summary Table for Chapter III

Evaluation of the Curriculum

Evaluation Problems →	Implementation of the Institution's Objectives	Recognition of the Instructional Needs of the Students	Recognition of Individual Differences Among Students	Articulation with Curricula of Other Educational Units	Meeting the Needs of Society
Available or Existing Measures and Techniques for Evaluation	1. Group study and discussion 2. Opinions and ratings 3. Follow up* 4. Comparison of graduation requirements with objectives 5. Study of attrition 6. Control group experimentation	1. Group study and discussion 2. Opinions and ratings 3. Follow up* 4. Comprehensive achievement tests 5. Study of attrition 6. Control group experimentation	1. Group study and discussion 2. Opinions and ratings 3. Comparison of graduation requirements with student abilities 4. Study of attrition which compares rates at different levels of ability 5. Study of distributions of scores on ability measures for separate classes	1. Group study and discussion 2. Opinions and ratings 3. Tabulation of duplication of topics covered 4. Follow up of graduates* 5. Experimental studies (e.g., early admission)	1. Group study and discussion 2. Opinions and ratings
Measurement Needs	1. Improved techniques for group study and discussion 2. Improved techniques for opinions and ratings 3. Objective measuring devices for use in experiments (?)	1. Improved techniques for group study and discussion 2. Improved techniques for opinions and ratings 3. Objective measuring devices	1. Improved techniques for group study and discussion 2. Improved techniques for opinions and ratings 3. Objective measuring devices	1. Improved techniques for group study and discussion 2. Improved techniques for opinions and ratings 3. Objective measuring devices	1. Improved techniques for group study and discussion 2. Improved techniques for opinions and ratings 3. Objective measuring devices 4. Criteria of success in important fields
Available Interpretive Data		1. Normative data on published tests	1. Normative data on published tests		
Needed Interpretive Data	1. Normative data for all measures 2. Normative data on reasons for drop-out	1. More extensive norms for published tests 2. Norms for all new measures 3. Normative data on reasons for drop-out	1. More extensive norms for published tests 2. Norms for all new measures	1. Norms for all measures 2. Tabulations of graduation and admission requirements	1. Norms for all measures 2. Norms for various occupational groups, etc. 3. Relations between grades and success in various fields

*See Summary Table for Chapter VIII, column 1.

PRECEDING PAGE BLANK- NOT FILMED

Chapter IV

THE FACULTY

Introductory Statement

The previous chapter stated that the curriculum is the heart of an educational institution. This is true in the sense that the teaching of knowledge and skills at advanced levels is a unique function of colleges and universities. But a curriculum without a competent faculty would not be an effective educational instrument. Indeed, the curriculum is very likely to be the product of the thinking of the faculty, and certainly the effectiveness of any curriculum depends to a considerable degree upon the faculty's performance in the classrooms and laboratories. The repute of an institution of higher education depends, perhaps, more upon the quality of the faculty than upon any other factor or group of factors. Unless an institution has a high percentage of first-rate scholar-teachers, it cannot claim to be in that group of colleges and universities entitled to the designation "distinguished."

How does one determine the quality of a faculty? How should one proceed to rank the faculties of various institutions? What are the important variables of faculty performance, and how does one go about the measurement of these variables? A great many general statements have been written about the qualities of the good teacher, but what are the fundamental, measurable "dimensions" of the college teacher? A list of these dimensions, with adequate techniques for measuring and evaluating them, would be an important contribution to faculty evaluation and college personnel administration.

Since the instructional effectiveness of the college teacher is so very important, and since instructional effectiveness is influenced by techniques and facilities as well as by the qualifications of the teacher himself, a separate chapter is devoted to the evaluation of instruction. The present chapter will be concerned with the qualifications of faculty members, their scholarship, personal attributes, and attitudes. The reader should bear in mind, however, that effectiveness in teaching is an important aspect of the faculty evaluation problem and is discussed in the next chapter.

Evaluation Problems

Qualifications of the faculty. It would seem logical that the first dimension of faculty evaluation should be that of personal and professional qualifications. One of these "qualifications" factors which has been seized upon at all educational levels as a means of improving the quality of the faculty is that of increased training. The elementary school has in recent years so increased its requirements that a B.A. degree is becoming quite common among elementary school teachers. Secondary schools in many states are requiring or are working toward requiring a master's degree. The general feeling among elementary and secondary school educators seems to be that increasing the training requirements of teachers has improved their quality. It seems quite logical to believe, therefore, that quality of teaching on the college level should also be associated with amount and type of training.

Several studies have been conducted on the relationship between the level of training of the faculty and the over-all quality of the institution (48 , pp. 59-65). An early study (1928) by Reeves and Russell among twenty-nine institutions belonging to the North Central Association showed a correlation of .43 between percentage of the faculty meeting North Central faculty training requirements and a judgment of institutional merit. The corresponding correlation in the Haggerty study was .36, and the correlation between percentage of the faculty having the Ph.D. and the over-all evaluation (Criterion VIII) of the 56 institutions in the study was .30.* Positive but lower correlations were found between percentage of the faculty having Masters' degrees and Criterion VIII and between years of study and Criterion VIII. While these correlations are fairly low, they do indicate that the amount of training of the faculty appears to be related to over-all institutional quality. Naturally the size of the correlation is directly related to the prevalence of advanced degrees among the faculty members of the colleges being studied. If all, or nearly all, members of the faculties had the Ph.D. degree, there would be little or no variation in the training variable, and hence the correlation would be near zero.

Another complicating variable in the "amount of training" dimension is that the requirements for teaching in various departments of a college or university vary a good deal. Professors of art, music, and creative writing, for example, sometimes lack even a B.A. degree. Their training may have been received in special institutes or under the direct tutelage of a master. Many of the outstanding teachers in these fields have had no graduate work at all. Unless this fact is borne in mind, serious misinterpretations can be made of data on the relationship between advanced degrees and the quality of the faculty.

Experience in teaching is another "qualifications" factor believed by both educators and laymen to be associated with quality of teaching, although at the lower levels (elementary and secondary school) there is evidence that experience beyond a certain minimum is not associated with superior teaching. It seems quite logical, however, that a faculty member's teaching ability should increase with experience. Positions of the greatest responsibility in the educational world, as well as in business and industry, generally go to the more experienced persons. Students who have just graduated from college or who have recently finished their graduate studies often feel that they are handicapped in competing for positions because of lack of experience. There would seem to be little question, therefore, about the appropriateness of experience as an aspect of the "qualifications" dimension to be used in evaluating the quality of a college faculty.

---

*A correlation of .30 is, in a sense, not very high. However, it does indicate a relationship between the variables. Since this report is concerned with group rather than individual performance, it would seem that correlations of this magnitude may have some degree of meaningfulness.

Haggerty, in his study of the faculty, correlated length of experience in teaching and administration in higher education of the staff with the over-all ratings of the institutions included in his study. This correlation was .29, roughly equal to that of the relationship between amount of training (number having Ph.D. degree) of the faculty and over-all institutional quality. The writers have recently been informed through personal correspondence of a study at the University of Minnesota in which students rated their instructors as effective teachers in direct proportion to rank; that is, full professors received the highest rating, associate professors next, and instructors were given the lowest ratings. Obviously these ratings also correlated positively with length of experience. Studies such as these leave unanswered many questions about the relationship between experience and success in teaching, but they do indicate that the variable of experience should be investigated. It seems to be one factor which should be included in any comprehensive appraisal of the quality of a college faculty.

Scholarly production. The faculty member in the better institutions is expected to contribute to his teaching field through the writing of books and articles as well as by presenting his subject in the classroom. While there may be considerable disagreement as to the degree of emphasis which should be given to research and writing, educators would be almost unanimous that scholarly productiveness is a very important attribute of a college or university faculty member and should be considered in the institution's appointment and promotion policy.

Performance as a productive scholar may not be so intangible as quality of teaching, but the evaluation of it is far from simple. Nearly all educators would agree that a mere enumeration of the books and articles written by faculty members would be an inadequate measure of scholarship. Ideally, any evaluation of scholarly production should be qualitative as well as quantitative. It seems safe to assume that in the actual appointment and promotion of faculty members a real effort is made to assess quality as well as quantity of writing. The survey studies which have been reported in the literature are primarily quantitative in nature and hence are lacking in completeness of evaluation.

In the North Central Association's study of faculty competence (48 , pp. 66-70) the staff members of the participating institutions were asked to list the number of books, articles, monographs, reviews, and non-professional publications produced during the "last five years" and also to give the total number of such publications produced during the professional lifetime of the individual. It was found that the highest correlation, .41, existed between total list of publications and Criterion VIII (over-all quality of the institution). Much as one may deplore the practice of simply counting publications, it is of considerable interest that a correlation of this magnitude should be found between faculty publication and institutional quality.

Associated with the factor of publications is that of membership in learned societies and attendance at annual meetings. Every major field of instruction in the college curriculum is represented by a professional organization, and it would seem likely that recognized scholars would hold membership in such organizations and participate in their activities. Haggerty (48 , p. 72) studied the significance of membership and participation in learned societies as related to over-all institutional quality and found the following correlations: attendance at meetings and over-all quality criterion, .50;

participation in programs and the criterion, .43; membership in professional societies and the criterion, .39. These relationships indicate quite clearly that scholarly production and membership and participation in professional or learned societies are correlated with the over-all quality of the institution. The evidence is that these factors constitute significant "evaluation dimensions" of faculty quality.

Attitudes toward scholarship, teaching, and students. Studies of human behavior, as well as general observations of the performance of people, testify to the importance of attitudes -- favorable or unfavorable, healthy or unhealthy -- in a wide variety of human activities and relationships, including political action, family life, social activities of all kinds, and effectiveness in vocational pursuits. College teaching is no exception.

The present writers are not aware of any studies which have been made of the relationship between the job attitudes of college faculty members and their over-all quality as scholar-teachers. There have been studies of job satisfaction in business, industry, and government, but very little of this sort of research has been done in academic fields. In the behavior of faculty members there are decided individual differences in attitudes towards scholarship, teaching, and students. Some faculty members regard research and writing as the highest form of activity; others would place teaching or counseling first. And there may be some who actually detest some of their duties as college faculty members. Our data on this question are as yet very limited, but measurements of these attitudes might be useful in the determination of the quality of a college faculty.

Personal attributes in relationship to colleagues. Personality is generally regarded as a very significant factor in vocational success. Indeed, some reports have suggested that it isn't so much the technical skill or training that counts in occupational performance as it is the personal qualifications of the individual. The statement is also made quite frequently that individuals lose their jobs or fail of promotion not because of incompetence but rather as a result of personality deficiencies. Studies of student evaluation of quality of instruction (Chapter V, p.105) also indicate that the personal attributes of the instructor are important in determining his teaching success.

College teachers have a great deal of independence in the planning of their courses, in teaching, in selecting and working on their particular field of research and scholarship, and in participating, or not participating, in extra-classroom activities. In spite of this, the work of a department or of the institution as a whole depends upon group effort. The admission of students, the formulation of requirements for the major, the construction of master's and Ph.D. comprehensive examinations, the scheduling of courses, and the handling of student problems all demand a considerable degree of teamwork. One can see, therefore, that the work of a department or of an entire institution depends in large part upon the ability of faculty members to get along together. A vitriolic, uncooperative individual, inconsiderate and disrespectful of his colleagues, can seriously interfere with the effectiveness of the department or institution. While individuality or personal idiosyncrasy is not undesirable, at what point does the idiosyncrasy of a faculty member impair the work of the unit to which he belongs? Does his presence on the staff help or hinder the recruitment of new faculty members and students? Are the actions and attitudes of the individual constructive or destructive? These questions, all related to the factor of interpersonal relationships, are very significant. Our information to date on this dimension of measuring the quality of a faculty is quite inadequate, even though it is generally agreed that personal qualifications are important in determining a faculty member's success. It should also be said, however, that the best colleges and universities encourage rather than discourage individuality in their faculty members. Personality becomes of concern only as it affects the quality of teaching and scholarship. What we need to know is why some personalities seem to promote the welfare of a department or college while others seem to hinder it.

Observance of professional ethics. College teachers, like most other major professional groups, have formulated, largely under the auspices of the American Association of University Professors, certain basic rules of professional conduct, which have been published in the journal of that association. Abbreviations or revisions of the AAUP statements are frequently found in college handbooks and other publications dealing with the rules and regulations of the particular institution. Statements on professional ethics typically are brief and emphasize primarily institution and colleague relationships and the topic of academic freedom. It is generally assumed that every faculty member is thoroughly informed as to rules of conduct in scholarship and research and in personal affairs.

Unfortunately, faculty members are not always careful to observe professional ethics. This is particularly true with respect to the personal criticism of colleagues (as opposed to their ideas and works), of students, and of the institution, especially at times and places which are inappropriate or even in bad taste. There are decided individual differences among faculty members in sensitivity to the appropriateness of time, place, and method of communication for valid criticisms of colleagues, students, administration, and governing boards. Healthy criticism is something to be sought and encouraged; how to get the right type of criticism at the right time and place is the problem. The measurement of the degree to which faculty members observe the accepted rules of professional conduct would have very interesting research and practical values. That these measurements would be related to individual and institutional quality would seem to be a tenable hypothesis.

Morale of the faculty. The visitor who comes to the college campus and talks to faculty members, students, and administrators and who observes the activities taking place on the campus will soon sense the tone of the institution and the morale of those connected with it. The morale of a group is difficult to define, but its significance is very real. A group which has the "right spirit" is going to overcome obstacles and accomplish things; on the other hand, the group with the "wrong spirit" is easily discouraged and fails to accomplish very much. What are some of the factors which influence the morale of a college faculty?

Haggerty (48 , pp. 127-177), in the North Central Association study of institutions of higher education, reports on the following factors which were believed to be associated with the effectiveness of the faculty: salary, years of service required to attain tenure, instructional load, methods of recruiting, faculty growth, leaves of absence, retirement, insurance, housing, and community life. The factors of salary, tenure, and instructional load could be reduced to arithmetical terms; in the case of the remaining factors ratings were obtained. Each of these measures was then related to the over-all quality of the institution (Criterion VIII), using the coefficient of contingency as the measure of relationship. The reported coefficients are as follows: salary, .42; tenure, .42; load, .36; recruiting, .55; growth, .47; leaves of absence, .54; retirement, .42; insurance, .45; housing, .32; community life, .41. From these coefficients it would appear that there are conditions of faculty service which are associated with the quality of the institution. It would seem fairly safe to assume that these factors are also associated with faculty quality, and one may well suspect that they would have a good deal to do with faculty morale.

Associated with conditions of service is a group of factors which Haggerty (48 , pp. 98-126) termed "faculty organization." This included such factors as quality of institutional leadership, faculty-student ratio, the assignment of faculty members to the fields for which they are trained, effectiveness of faculty organization, nature of faculty meetings, and use of faculty committees. Most of these factors are of such a nature that ratings had to be used in assessing them. When all of the separate factors of faculty organization were integrated into a single factor, and this index correlated with over-all institutional quality, it was found that the coefficient of correlation was .67. This coefficient, as reported, is actually greater than that obtained for the relationship between faculty competence and over-all institutional quality. It should be remembered, of course, that the rating of the factors making up "faculty organization" may possibly have been influenced by the judges' estimate of the quality of the institution as a whole. Despite this fact, it would appear that this thing which Haggerty calls faculty organization is positively related to institutional quality. Again, one would suspect that the factors which make up faculty organization might have a considerable bearing upon faculty morale.

To the knowledge of the writers, no reported studies present quantitative data on the status or level of faculty morale at various institutions. A Faculty Morale Scale, prepared by a chapter of the American Association of University Professors, has recently been placed on the market by Psychometric Affiliates, Chicago, Illinois. Actually the scale consists of a list of 34 institutional characteristics such as "adequacy of salaries," "sabbatical policy," and "faculty attitude toward teaching," each of which is rated on a five-point scale, ranging from "very good" to "very poor." Percentile scores, based upon a very limited number of faculty respondents, are reported in the manual. The list of items included in the scale might possibly prove useful in studying the status of morale on a particular campus; the data furnished in the manual would be of doubtful usefulness.

Admittedly, the factor of morale is illusive, but the indirect evidence suggests that there is something in this dimension which is related to institutional quality. It would seem, therefore, that in measuring the quality of a faculty, it would be important to have data on the level of faculty morale.

Contribution to school and community activities. The central functions of the college professor are teaching and research, and his success should be judged primarily by his performance in these areas. Nevertheless, it should be remembered that a faculty member is also a citizen of the college community and of the city in which he resides. In these days of increasingly complex social, economic, and scientific problems, it would seem that a considerable number of faculty members should concern themselves with these problems and with the interpretation of them to the public. It is also essential that the college or university itself should be described and explained for the benefit of those who may be unfamiliar with the basic functions of higher education and the role which it plays in our society. There is, in other words, an important public relations task to be performed, and college faculty members as well as administrative officers should participate in it.

The education of the student is a complex task and is not limited to the activities which go on in the classroom and laboratory. The organized extra-curricular activities and the informal contacts between faculty and students are also very important. Advising or counseling is another illustration of an extra-classroom activity which is a significant part of the student's total educational experience. The faculty member who takes seriously his responsibilities in all these areas is, other things being equal, contributing more to his institution than the one who avoids such participation.

Faculty members are sometimes accused of not taking a deep interest in the community in which they reside, or, at least, of having interests which seem to differ from those of other citizens in the community. Again, it is essential that the college or university faculty bear its proper share in civic and community affairs and that faculty members work for the betterment of the entire community as well as of the institution with which they are affiliated. Faculty participation in community affairs should also contribute substantially to the professor's own growth. The sociologist who works with civic organizations, the chemist who consults with industry, the economist who associates with business and labor leaders -- all will probably benefit from these experiences and gather a wealth of practical illustrations to enliven the classroom.

In brief, the extent to which a faculty participates in school and community activities would seem to be another dimension which should be measured in the evaluation process.

Techniques of Evaluation

Biographical data sheets. Data presented in the preceding section indicate that such factors as training, experience, and scholarly production of the faculty are associated with institutional quality. If information on these questions is needed for evaluation purposes, how can it be obtained? The answer might be the biographical data blank, which usually asks for essential personal history data, schools attended and degrees earned, positions held, courses taught, scholarly books and articles written, program of scholarly activity, membership in professional organizations, meetings attended and offices held, prizes and honors won. Some colleges require that these data blanks be not only filled out when the person is first appointed, but also kept up to date, especially with respect to scholarly production, meetings attended, and honors won. While there may be some unreliability in the data found on personal data blanks, they are on the whole free from error when proper directions are given for filling them out. Such errors as do occur are usually quite unintentional.

While the biographical data blank is very useful in the collecting of essential information, it in itself is not a measuring instrument. The armed services, business firms, and various government agencies have succeeded in developing scoring devices which can be used to evaluate the data found on biographical inventories, but such techniques have been used very little if at all in higher education. The North Central Association, in its study of higher institutions, published percentile rankings (48 , p. 85) for the fifty-six institutions included in its investigation, with respect to advanced degrees held by the faculty, years of experience, books and articles written, membership in learned societies, attendance at meetings and participation in programs of learned societies. More recent percentile rankings for schools in the North Central Association have been published in the North Central Quarterly (62). Since it has been shown that each of these factors is related to institutional quality, these percentile tables may be of some usefulness in evaluating the quality of the faculty at a particular institution.

To some it may appear that such factors as percentage of the faculty holding advanced degrees, years of experience, average number of books and articles written, membership in learned societies, and participation in learned society meetings may not be very significant in the measurement of faculty quality. Standing by themselves they may not mean very much, but they are symbolic of something which is important. Faculty members who have earned advanced degrees, who share with their colleagues new ideas, and who seek out the company of their fellows at other institutions are "alive" in their profession, and are likely to be effective teachers of young people. The evidence from the North Central Study is that faculty members ranking high on these factors are found at the better institutions. The norms which one can obtain from biographical data blanks may not be particularly useful in the evaluation of individual faculty members, but they would appear to have merit in the evaluation of faculties at various institutions.

Rating scales. Since the evaluation of a college faculty is a complex problem involving many variables, it is not surprising that the most widely used evaluation technique is that of rating. Ratings may be made by colleagues, superiors, outsiders, or the individual himself, and may vary from very informal comments to highly formal evaluations involving the use of rating scales and thorough, statistical analyses of results.

Ratings are most likely to be made at the time of appointment and when the faculty member is being considered for promotion. Some institutions, for example, require that all new appointments and reappointments as well as promotions in rank be endorsed by a majority of the membership of the department. The University of Washington faculty rating program described by Guthrie (25) is probably one of the most extensive in terms of the number of persons rated and the use made of the results.

It is unquestionably true that colleagues know a good deal about one another and that their ratings can be useful in faculty evaluation. At the same time it should be said that the rating of colleagues whom one knows well is a delicate matter. An administrator or faculty committee can obtain informal comments about faculty members considered for promotion, but resistance is likely to be encountered the moment the rating procedure is made formal. In a sense, no better form of evaluation can be obtained than that of rating by colleagues, but the practical difficulties in obtaining the ratings are considerable.

The use of outside ratings in the evaluation of administrative officers is illustrated in the North Central Association study

(48 , p. 101) of institutions of higher education. The rating scale used in the study included ten factors, each of which was rated on a scale ranging from 1 to 9. The total scores on the scale, therefore, could range from 10 to 90. The ratings (made only of college presidents) showed considerable variation and seemed to give a fairly good picture of the institution over which the president presided. It was felt, however, that in view of the personal nature of these ratings they should not be used in college accrediting. Despite the fact that the ratings were not used in the study, the discussion of the results would seem to indicate that this technique does have promising possibilities.

Clark (11) has described an interesting procedure for the identification of a "significant contributor" group in psychology, using what is basically a counting of publications and a rating technique. The first step in the process was to count the number of Psychological Abstract entries for psychologists grouped by the periods in which they were awarded their Ph.D.'s, as follows: (1) 1930-1934, (2) 1935-1939, and (3) 1940-1944. The top 150 names in each of these three lists were labelled "high producers." A group of 22 competent judges was then asked to review this list and to add names which in their judgment should be included. The expanded lists were labelled "highly visible psychologists" and totalled 624 for the three groups (1930-1934; 1935-1939; 1940-1944).

The lists of "highly visible psychologists" were then sent to all whose names appeared in the lists, with instructions to nominate not more than 25 who in their opinion had made the most significant contributions to psychology as a science and to select the best 10 persons in their particular areas of psychology (five major areas were designated). The final list was selected as follows: First, the top five persons in each of the five areas of psychology, in terms of the number of votes received as being "among the best 10 in the area," were selected as significant contributors. This gave a total of 25 persons for each of the three five-year periods. The other 25 persons (for each period) were selected as those receiving the largest number of votes as being "among the best 25 in psychology" who had not already been selected in the first step. The final list, then, consisted of 150 members of the American Psychological Association designated as "significant contributors."

After the criterion group had been obtained, an effort was made to identify objective factors which were correlated with membership in the list and which might serve, individually or collectively, to identify persons included in the group. The four factors which in combination gave the best prediction of membership in the list were as follows: (1) 1950-53 Psychological Abstracts counts; (2) Annual Review (of psychology) citations; (3) journal citation counts (how many times a man's work is cited in the articles appearing in the best known psychological journals); (4) offices held in the American Psychological Association. The multiple correlation of these four factors with number of votes received was .82. This correlation seems to indicate that knowledge of publications and offices held in the APA would give a fairly solid basis for determining whether a person would be nominated to the "significant contributor" group of psychologists.

The technique described by Clark could readily be adapted to most of the major fields of learning. It would be very interesting to have such information and to determine the identity of the significant contributors in the various colleges and universities of the country. The American Psychological Association is carrying out an intensive study of the group of 150 significant contributors. The factors associated with notable achievement in psychology will, no doubt, be of considerable interest to other fields. It is quite likely, too, that the completed study will contribute to the general problem of evaluating the quality of a faculty. This type of study, of course, does not evaluate teaching, since its entire concern is with scholarly production. But, as shown by Haggerty, the scholarly production of a faculty is definitely related to the total quality of the institution.

One or two of the shortcomings of the rating method, pointed out many times in books on educational and psychological measurement, seem to have special significance in the case of faculty ratings, particularly ratings made by colleagues. There is, first of all, the "halo" effect. Since departmental faculties are fairly intimate "family groups," it is difficult for the rater to isolate the particular factor being rated and to be objective about it. As with all ratings, this usually results in higher ratings than may actually be deserved. This "family atmosphere" of a department creates the further complication that rating may take on an aspect of politics, thereby reducing objectivity. If everyone is "for" everybody else, the ratings will be too high; if there is something of a family quarrel, personal differences may easily over-ride good judgment. To state it another way, rating by colleagues within a department is a very delicate affair and unless it is very carefully managed may result in an unhappy situation. Perhaps this more than any other reason accounts for the fact that rating scales are seldom used by colleagues in the evaluation of their associates on a college faculty.

A second shortcoming of ratings by colleagues is that quite frequently the experience of faculty members is too limited to permit them to give valid ratings of a colleague -- at least in any detailed sense. Since the work of college faculty members is highly specialized, faculty members who are not in the same specialty are naturally reluctant to evaluate a colleague with whose field they are not intimately acquainted. Perhaps college faculty members underestimate their rating ability, but in a practical situation one encounters this difficulty.*

Scholastic records and test scores. Scholastic records and educational achievement test scores are frequently considered in the appointment of new faculty members who have just completed their graduate studies. Such records and scores might also be useful in the evaluation of a college faculty. It seems reasonable, generally speaking, that students with the highest scholastic records are likely to be the best faculty members. The experience on the elementary and secondary levels has been that college grades are moderately related to teaching success -- usually as evaluated by superiors. The relationship between academic grades and success in college teaching has not received extensive study. One of the chief reasons for this is that systematic evaluation of teaching success on the college level is considerably less common than it is on the elementary and secondary school level. In other words, the criterion data needed to evaluate the scholastic record as a predictor of teaching success on the college level are difficult to obtain. Many placement offices receive letters from college administrators commenting on the success of recently appointed college teachers, but these are, for the most part, not studied systematically. As will be indicated in the next chapter, systematic evaluation of teaching on the campus is not routine. Since grading standards vary somewhat from institution to institution, it would be helpful in the evaluation of a college faculty to have test scores in subject-matter achievement tests, modelled after the Graduate Record Examination but of an even more advanced character. The transcript of grades and the test scores could then be studied together as a means of assessing the quality of a college faculty.

---

*The reader may wonder why student ratings are not discussed as a device for evaluating faculty members. The writers believe that students are in a position to evaluate one phase of faculty competence, namely, teaching effectiveness. Student ratings are, therefore, discussed in the next chapter.

Letters of recommendation and personal interviews. Along with the scholastic record of a prospective faculty member, several letters of recommendation are likely to be submitted. If the placement office has not provided such letters, or if there is reason to want to supplement them, it is rather general practice to write for information to former teachers of the person in question or to colleagues in the field, or both. And after scholastic records and letters of recommendation have been assembled it is general practice to ask the prospective faculty member to come for a personal interview. What are the merits of these two evaluative devices as a means of assessing the quality of a college faculty?

From the standpoint of statistical data there actually is very little evidence that letters of recommendation and personal interviews are very discriminating in the information which they provide. Seldom do letters of recommendation convey any negative information! Also, letters of recommendation are frequently written in such general terms that they are not helpful in a particular situation. The personal interview is notorious for its lack of validity as a selection device in business, government, and industry, and there is no evidence that it is any more valid in selecting college teachers. The main purpose served by the interview is to give some indication as to whether or not the prospective faculty member will fit in as a member of the group. Even this is difficult to judge in a short time, but there is at least some chance that this type of evaluation may result from a personal visit. While considerably over-rated, there seems to be little doubt that both the letter of recommendation and the personal interview will continue to be used in the selection of new staff members. The accumulated letters of recommendation and the records of personal interviews might conceivably have some degree of usefulness in the evaluation of the quality of a college faculty.

Attitude scales. The attitudes of faculty members toward scholarship, teaching, and students could conceivably be measured by means of attitude scales. Studies in social psychology show that individual differences in attitudes towards activities and institutions, as well as changes in attitudes, can be measured by attitude scales. The generalized type of attitude scale developed by Remmers (57) might have some possible usefulness in the measurement of faculty attitudes towards various aspects of their jobs. To the knowledge of the writers, attitude scales have not been thus employed. It seems quite likely that any attempt to use attitude scales with college faculty members would be met with a good deal of skepticism and perhaps hostility. The best use of attitude scales in the evaluation of college faculties at present would be on an experimental basis. A study of attitudes would seem particularly appropriate for the younger members of the faculty and for graduate students training to become college teachers.

Amount of turnover. Rate of turnover is generally regarded as an indication of the healthiness of the personnel situation in a particular organization, whether it be government, business, or education. In general it would seem that colleges with the best faculties would have the least turnover and those with the weakest faculties the greatest turnover. Naturally, the percentage of turnover should never be taken at face value. It is conceivable, for example, that the faculty at a particular institution might be of such poor quality that no other institution would want to offer a position to anyone on its staff. While the turnover in such a college might be very small, it would obviously be a serious error to regard the low turnover rate as a mark of strength. On the other hand, a school might have a fairly high turnover because it attracts very able young men and has not enough openings on the senior staff to be able to hold them. In other words, the level of academic rank at which turnover occurs is very significant. Thus, while the rate of turnover may be an indication of quality, there are definite exceptions to the rule.

The relationship between turnover and institutional quality has not been studied extensively. Haggerty (48 , p. 66) found a correlation of .29 between average years of experience in college teaching and administration of the faculty and over-all quality of the institution. The average length of the term of service of members of the faculty varied a good deal from institution to institution. Haggerty reports that in eleven of the 56 institutions studied, the average experience of the faculty was five years or less; in one it was less than three years, and in no institution was it as high as thirteen years.

The average for all fifty-six institutions was 8.4 years. The correlation coefficient reported by Haggerty, while not a direct measure of turnover, would seem to suggest that turnover is positively correlated with institutional quality.

Evaluation Needs

Better evaluation of quality of preparation. Every administrator who has ever been responsible for the appointment of new faculty members who have just completed their graduate work has yearned for a better means of assessing a candidate's potentialities for college teaching. As indicated earlier, the scholastic record, test scores, letters of recommendation, and interviews leave much to be desired as predictors of teaching success and scholarly promise. Two additional bits of information would seem especially desirable. First, beside his scholastic record, how high are the candidate's originality as a scholar, the quality of the papers written by him, his energy, resourcefulness, and imagination? All these may be touched upon in letters of recommendation, but they are likely to be evaluated on the favorable side. How does the individual in question really compare with other graduate students? If each instructor of graduate students were to keep such information and report it along with grades it would contribute greatly to the evaluation of the potential college teacher. There would be problems involved in reporting this type of information to potential employers, but it would be good to have such data, at least for research purposes.

A second type of very useful information would be evaluations of success in teaching. Nearly all prospective college teachers are employed at one time or another as graduate assistants or part-time instructors. The teaching done by graduate students is, or should be, supervised by senior faculty members; hence it would not take a great deal of extra effort to provide evaluations of success in teaching. Undergraduate students who plan to enter public-school teaching are required to take a course in practice teaching; they receive a grade in the course. It seems rather unlikely at present that prospective college teachers will be required to take a course in practice teaching in which they would be graded just as they are in their academic subjects. Nevertheless, many of them do "practice teach" while enrolled in the graduate college. A careful evaluation of this teaching could well serve in lieu of a practice-teaching course to provide information on a point which is now covered inadequately in the records of young college teachers. If the data pertaining to quality of preparation could be thus improved, one would be able to make better use of this type of information in the evaluation of a college faculty.*

Better ratings of faculty qualifications. The evaluation of a college faculty is a complex task. Useful information can be obtained by means of biographical data sheets, but to a considerable extent the evaluation must depend on the application of human judgment. A crucial question then is: How can this judgment process be improved? The difficulties generally encountered in the application of judgment have been discussed in previous chapters. What needs for improvement exist particularly in the field of faculty evaluation?

In the previous section it was stated that ratings can be made by colleagues, superiors, outsiders, and the self. It is quite clear, however, that these different raters cannot make the same types of ratings. Colleagues, for example, may be in the best position to judge scholarship; superiors may be best able to determine how well a particular faculty member serves the institution as a whole; outsiders can usually best assess a particular man's contribution to his professional field; and, perhaps, the faculty member himself is the best judge of his satisfactions with his work. The role of students in the evaluation of college teaching will be discussed in the next chapter.

---

*The recommendations in this section are in general accord with those recently published by a Committee of Fifteen (67 , pp. 19-22).

It would seem that a study could profitably be made to determine more precisely just what aspect of a faculty member's performance can best be evaluated by each type of judge. Unless a judge has the information necessary for judgment, his evaluation will not be of much value. There seems to be a need, therefore, to determine how the total rating task can best be distributed among a group of judges.

A good many colleges and universities have instituted faculty rating programs on a hit-or-miss basis. What seems to be needed now is to institute a number of systematic rating programs in a variety of institutions. Ideally such ratings should be made by colleagues, supervisors, outside judges, students, and perhaps the faculty members themselves. Systematic rating programs would make it possible to chart changes in rating over a period of time and to correlate various factors with these ratings. At present the available data on ratings are fragmentary. If data could be obtained for a particular faculty over a five- or ten-year period, a good deal could be learned about rating as a method of faculty evaluation. Such data may be available at a few institutions such as the University of Washington, but "longitudinal" rating programs are the exception rather than the rule.

The study of psychologists by Clark (11), described in the preceding section, illustrates how the task of evaluation can be accomplished using the judgments of colleagues at other institutions plus the record of publications. It would seem that this type of study could well be duplicated with many other groups. If all the fields represented by departments in various colleges and universities were to conduct such studies, one would, at least, be in a position to determine how many well-known people there are on the faculties of various institutions. At the present time reference is sometimes made to the number of faculty members in Who's Who in America or American Men of Science. The type of study being made of psychologists would be far superior to any listings which are now available. To be sure, such studies would not give a complete evaluation of a whole faculty, but they would provide a "highly visible" faculty index which might well be correlated with the quality of the entire faculty.

Finally, and perhaps most importantly, a study should be made to determine how a faculty rating program can best be carried out. At the present time a great many faculty members are skeptical as to the value of rating. Some departments and some institutions are so suspicious of rating systems that they would be unwilling to experiment with them -- let alone to place confidence in the results. Faculty members are well aware of the fact that some form of evaluation is inescapable, but, so far, not many would support a formal rating program. This problem is worthy of study, and until it is solved, ratings of faculty performance will not find a prominent place on the college campus.

Better norms for evaluation of biographical data. The North Central Association study of the evaluation of colleges and universities, cited frequently in this chapter, provided a limited number of normative data which could be used for the evaluation of such factors as education, experience, scholarly production, and membership in learned societies. More recent normative data on these factors have been published in the journal of the North Central Association.*

The chief need at present is for much more refined norms for these factors, norms by type of college -- junior college, liberal arts college, teachers college, individual colleges of universities -- and, for each type of college, norms by academic ranks and fields of knowledge.

---

*A study of the baccalaureate origins of the faculties of liberal arts colleges, teachers colleges, and junior colleges is being made under the direction of Dean Frank R. Kille, Carleton College, Northfield, Minnesota. This study will provide information with respect to age and degrees of faculty members in each subject-matter field for the three types of colleges mentioned above. The study will be completed in 1957 and will provide the most extensive data yet made available on the baccalaureate origins of college faculty members.

The North Central data were gathered for the purposes of overall institutional evaluation, hence the more refined data were not necessary, but if individual colleges are to use such data for the improvement of their faculties, the more detailed and refined information is essential. It seems quite logical, for example, that quality of teaching rather than scholarly production should be emphasized in junior colleges; that faculty members of the rank of instructor would not have as extensive lists of publications as do full professors; that faculty members in the physical sciences in general would probably have more publications than those in mathematics. All these items are very important in evaluating the quality of faculty members. Such normative data as are here suggested would have to be applied with some caution, but the more refined they could be made the more useful they would become. One can hardly think of any objective data which would be more helpful in the evaluation of college faculties than would these tables of norms. If they could be gathered in such a way as to include a thoroughly representative sample for each college type and could then be distributed to all colleges for informal use, much good could be accomplished.

Better assessment of attitudes. At present the assessment of the attitudes of faculty members and potential faculty members toward scholarship, teaching, and students is done, if at all, by general observation, often on a very informal or casual basis. Granted the importance of attitudes, can any techniques be suggested which would increase the validity and reliability of these observations and reports?

The techniques which immediately suggest themselves are attitude scales, interest inventories, and ratings. Despite the fact that attitude scales and interest inventories have been used for a good many years, the scores obtained from these instruments have not been especially useful in personnel administration. Since attitude scales and interest inventories are of questionable validity in the evaluation of personnel, it would not seem proper to recommend that they be used in faculty evaluation except on an experimental basis. As mentioned in the preceding section, ratings are also of limited usefulness in a faculty evaluation program -- particularly for the measurement of such intangible factors as attitudes.

The best hope would seem to be in the improvement of rating graduate students while they are training to become college teachers and during their early years as college teachers. If such ratings could be made systematically over a period of two or three years and then summarized they should prove of considerable interest in the evaluation of a college faculty, at least for research purposes.

Better evaluation of personal attributes. Personnel research in government, the armed forces, business, and industry has demonstrated the importance of personal attributes in bringing about maximum efficiency of the work of groups. Studies of this sort are almost unknown among faculty and professional groups in higher education. It seems likely that studies of personal attributes, particularly attributes affecting working with others, would contribute very considerably to our understanding of what makes for effectiveness or ineffectiveness in college teaching and research.

Studies of the personal attributes of faculty members would be difficult to carry out, particularly if one attempted to identify cooperative and uncooperative members of a group. It is possible that studies of the sort conducted by Roe (61) could be made of the memberships of entire departments, thus providing extensive data on personal attributes which could then be correlated with peer ratings of the cooperativeness of the members of the department or group. It is unlikely that this type of study would be cordially received in many institutions, but it might be worth trying.

A second possibility would be to assemble extensive personal and intellectual data on graduate students, much as was done in the Kelley and Fiske study of clinical psychologists (36), and then follow these students in their teaching and research careers. This type of longitudinal study would require time but over a ten-year period could provide data which would probably be useful in the assessment of the personal attributes of faculty members.

Another approach to the study of personal attributes is suggested in the leadership studies conducted under the sponsorship of Ohio State University (19). While it does not necessarily follow that a leader has the most desirable personal attributes, it is generally true that the leader has a measure of acceptance by the group. For example, a department chairman who is endorsed for his position by the membership of the department as well as by the administration of the institution is likely to have personal attributes which are approved by the group. If studies could be made of those recommended for chairmanships and the results compared with those obtained by a study of those rejected for chairmanships, perhaps some leads could be obtained to the personal attributes which make for good teamwork within a department. While the difficulties to be overcome are very great, better evaluation of the personal attributes of faculty members and prospective faculty members is very much needed.

Evaluation of the faculty as members of a group. An alternative to obtaining measures of cooperativeness on individual faculty members is to evaluate various faculty groups, for example, the members of one department or of one committee, as a team.

Hemphill (28) has reported the development of a 150-item questionnaire which shows considerable promise for use in describing faculty group characteristics.

Each member of the group, answering the questions independently, responds by indicating the extent to which the items apply to the group under consideration. The scoring results in group measures along 13 different dimensions.

Although the questionnaire is still in experimental form, experience in applying it to various groups has resulted in large differences in the expected direction. Further development of this questionnaire and studies showing which of the group characteristics are indicative of desirable faculty situations would seem well worth while.

A second possibility for describing faculty cooperation is to obtain a measure of "team coordination." This measure, described by Hemphill in a personal communication, requires the development of a hypothetical but realistic problem, the solution to which might involve all the members of the group. The problem is organized to include several steps of specific action, and the members of the group are asked, individually, to indicate who would take what action and when. The score is a measure of agreement among the members as to how the problem would be handled by them as a staff.

A preliminary application of this technique to an Air Force situation (31) was particularly encouraging. It would be of considerable interest to study the usefulness of the technique on an experimental basis in a number of college or university departments.

Improved letters of recommendation and interviews. Since letters of recommendation and interviews are used so extensively in the appointment of college faculty members, it would seem that the letters of recommendation and summaries of interviews existing in college files should be of value in the assessment of the quality of a college faculty. It would be of considerable interest to know whether the letters of recommendation and interview data concerning a good faculty differ in any significant way from those of a faculty of lower quality. Even though much use is made of letters of recommendation and interview data in judging the qualifications of individual faculty members, there are few, if any, data on the usefulness of these techniques in assessing the quality of the faculty of a department or college. It would seem studies of these techniques should be made.

It would seem that studies of the usefulness of letters of recommendation and interviews in the evaluation of college faculties should be instituted in a number of different institutions. While not all colleges would have the necessary data in their files, probably a number of colleges or individual departments could supply the necessary letters of recommendation and interview data concerning members of their faculties. These should then be rated by two or more judges. Reliability data should be computed, and, where possible, the ratings should be correlated with other indices of quality, such as scholarly production, teaching effectiveness, and service to the institution. At present most colleges and universities would have to base a judgment on their impressions; very few have made systematic studies of the usefulness of letters of recommendation and interviews. Through the data obtained in such a study, it should be possible to determine whether or not letters of recommendation and interview data are of any real value in assessing the quality of a college faculty.

Monitoring of the teaching process. A frequently encountered statement in letters of recommendation is "I haven't actually had an opportunity to observe the teaching of Dr. X. I have never visited his classes, hence my remarks are based entirely on second-hand information." It is unfortunately true that these statements describe quite accurately the situation which exists on many college campuses. Careful monitoring of the teaching process would seem to constitute an effective method for judging the quality of a faculty, but to date this evaluation device has not had wide use. The reasons for this situation are explained more fully in the next chapter.

Since it is not easy to monitor the teaching of mature, experienced faculty members, it would seem that one positive first step which could be taken by all graduate schools would be that of having senior staff members visit the classes of all teaching assistants and report on the quality of the teaching observed. Such observations would provide a measure of the quality of the teaching done by the junior members of the staff and might lead to considerable improvement of teaching in the department. Some departments already have almost this program, with close supervision given the teaching assistants by senior staff members. It would be highly desirable to have this practice adopted by all departments engaged in the training of college teachers, and thus to accustom the prospective teacher to a program of observation by superiors. It seems incredible, but it is nevertheless true, that many young teachers have never had their teaching observed by the senior staff. This situation could be corrected without great difficulty and such correction would do much to encourage the use of monitoring of the teaching process as a technique for faculty evaluation.

It should be said that some colleges have developed a tradition of easy intervisitation among all members of the staff, young instructors sitting in on the classes of more experienced professors and of one another. Such visits are of value to both visitor and host, particularly if there is extensive discussion afterward, and the tension of the monitoring type of visit is largely avoided. Because of the frequency and naturalness of the visits, a truer example of the teacher's prowess is probably exhibited and any subsequent assessment of colleagues' judgments concerning teaching qualities is made more valid.

Summary of Evaluation Needs

The preceding section has enumerated some of the major needs which exist in the field of faculty evaluation. In the accompanying table these needs are presented in summarized form. Which of these needs stand out most distinctly?

First, there is obviously a need for the improvement of evaluation devices -- biographical data blanks, rating scales, educational achievement tests for students finishing graduate work, inventories and tests for the assessment of personality attributes, letters of recommendation and interviews. The need for better measuring instruments is common to all institutions and organizations which have attempted to make more scientific their personnel practices and procedures. A great deal of progress has been made in the past 25 years in this whole field, but much remains to be done. All those interested in personnel administration, whether in higher education or in other fields of endeavor, will continue to work for the improvement of these instruments.

Second, there is a need in higher education for better norms to evaluate performance, particularly the scholarly production or reputation of faculty members. The North Central Association has made a good beginning on this problem, but much more should be done by colleges of different types, by departments, and by years of experience. At the present time, each college or university is practically on its own in the establishment of policies governing the appointment and promotion of the faculty and professional staff. Administrators and faculty committees charged with responsibility for these matters could improve their procedures if they could know more definitely what quality and quantity of performance can be expected of instructors, assistant professors, associate professors, and full professors located in departments and institutions of a particular type. A good many letters are written each year by individual institutions seeking to find out what is being done in other places in regard to faculty evaluation. The availability of norms of performance would not provide direct answers for the problems existing on a particular campus, but they would be of material help.

Third, it would be interesting and illuminating to have studies of college faculty members in which the characteristics of superior performers would be contrasted with those of average or below-average performers. The general approach of identifying significant contributors, used by the American Psychological Association, could be adapted for this purpose. Actually a good deal is known on every campus about the total contribution to the institution of those faculty members who have been on the campus for a period of five or more years. An investigation of this sort would have to be conducted with considerable tact; on small college campuses it probably would not be feasible at all. In the field of student personnel work a good deal has been learned about students by contrasting superior and inferior students. If the problems of carrying it out could be overcome, it would seem that such studies would also be fruitful in the area of faculty evaluation.

Fourth, it would seem that there is a need to know more about the interests, attitudes, values, and motivations of college faculty members. We know, for example, that some very superior teachers and scholars remain all their lives at institutions having relatively poor salary scales and lacking many of the outward qualities of superior institutions. What is love of teaching and scholarship? More information about these matters would be very helpful in the guidance of students planning to enter the field of college teaching and in explaining the successes and failures of individual teachers or groups of teachers. The studies made by Roe (61) of outstanding scientists and the American Psychological Association study directed by Clark (11) are suggestive of the approaches which might be used. The type of research to be done in this field is extremely complex and might well yield very little which would be of real help in the evaluation of college faculties. Nevertheless, the problem exists and should constitute a challenge to those who would seek to know more about the "why" of superior and inferior performance in college teaching and research.

Fifth, we need to know more about the different weights which might be given to faculty qualifications in different types of institutions or even in a single institution. The criticism is sometimes made, for example, that universities place too exclusive emphasis on productive scholarship and too little on teaching. We need to find out, more definitely than we know now, whether research or productive scholarship is necessary for good teaching at various levels. Perhaps educational institutions have need for three groups of faculty members: a) good teachers who do little or no research of the conventional sort; b) researchers who do little teaching, particularly of undergraduates, and c) the limited few who do both outstandingly well. In any case, some colleges possibly should select faculty members on one weighting of factors and universities on another weighting. If such weightings could be devised, the task of evaluating the faculties of different types of colleges and universities would be greatly expedited.

PRECEDING PAGE BLANK— NOT FILMED

Summary Table for Chapter IV

Evaluation of the Faculty

Evaluation Problems →	Qualifications of the Faculty	Scholarly Production	Attitudes toward Scholarship, Teaching, and Students	Personal Attributes in Relationship to Colleagues	Observations of Professional Ethics	Morale of the Faculty	Contribution to School and Community Activities
Available or Existing Measures and Techniques for Evaluation	1. Frequency counts of number of degrees 2. Distributions of number of years experience 3. Ratings of credentials 4. Counts of membership in and certification by professional organizations	1. Counts of publications 2. Ratings of the quality of publications 3. Ratings by colleagues	1. Generalized attitude scales, e.g., Remmers (57) 2. Frequency counts of the number of publications concerned with teaching	1. Personality inventories		1. Turnover rate among faculty members by rank	1. Distributions of hours spent in nonacademic campus activities 2. Frequency counts of the number of community activities participated in
Measurement Needs	1. Advanced subject-matter tests for Ph.D. candidate level 2. Record of teaching effectiveness as a graduate assistant	1. Improved techniques for estimating the quality of publications	1. Specific attitude scales (self) 2. Rating scales for use by colleagues 3. Objective ways of estimating degree of concern about teaching	1. Rating scales for use by colleagues 2. Improved personality inventories	1. Rating scales for use by colleagues	1. Job satisfaction scales 2. Rating scales for use by outside observers	1. Rating scales for check-lists to measure both extent and value of contributions to school and community
Available Interpretive Data	1. North Central Ass'n frequency counts of degrees	1. North Central Ass'n frequency counts of number of publications					
Needed Interpretive Data	1. Normative data for all measures 2. Relationship between specific biographical items and success*	1. Normative data for all measures	1. Normative data for all measures 2. Relationship between attitude and success as a faculty member	1. Improved norms for all measures 2. Relationship between personality traits and success as a faculty member	1. Norms, especially by departments within an institution	1. Normative data for all measures	1. Normative data for all measures

*See Summary Table for Chapter V.

PRECEDING PAGE BLANK- NOT FILMED

Chapter V

INSTRUCTIONAL EFFECTIVENESS

Introductory Statement

Education is concerned with the changing of behavior in terms of the acquisition of new skills, the learning of facts, the understanding of certain fundamental concepts and ideas, the development of attitudes and appreciations, and the acquisition of modes of thought and problem-solving skills. These changes in behavior are brought about through the curriculum, the student personnel services, the total impact of the college environment, the playing field, and most important of all, the classroom and laboratory. The unique function of the classroom or laboratory is the bringing together of the curriculum, the teacher, and the student. This chapter will be concerned with the effectiveness of the educational process in the classroom and its impact upon students as measured by changed behavior. How can one assess the effectiveness of the instruction which takes place in the classrooms of a particular institution? What are the dimensions of measurement of teaching effectiveness, and how can one proceed to quantify or otherwise assess them?

Evaluation Problems

Measurement of student performance. The real and perhaps the only test of instructional effectiveness is the extent to which changes in behavior are observed in the students who sit in classes or work in laboratories. Regardless of the personality of the teacher, the attractiveness of the classroom, or the quality of the curricular materials, if students do not learn to read better as a result of reading instruction, if they do not gain a knowledge of mechanics as a result of studying introductory physics, if the study of mathematics does not lead to a better use of quantitative methods, and if the study of the fine arts does not lead to the better appreciation of the arts, one may well question whether the educational program is producing worthwhile results. Evaluation of the curriculum, the student body, the personnel program, and the faculty itself does not mean very much unless it can be shown that important changes in the desired direction are produced in the lives of students. To be sure, many institutions perform functions other than that of teaching, hence it is entirely valid to use other criteria in the evaluation of their quality or usefulness to society. In the area of teaching, however, the critical test would seem to be: How much do the students learn in particular subjects and in their courses of study as a whole?

If it is true that student performance is the best measure of instructional effectiveness, one may ask why it has not been more widely accepted as the most valid test of the quality of an educational institution. The answer would seem to be that there has been considerable uncertainty how far all the important outcomes of instruction can be assessed by the available measuring instruments. If all facets of student performance could be measured satisfactorily, most educators would probably agree that performance is the best measure of the effectiveness of an educational program.

The term "student performance" is broad in scope. What particular aspects of performance are of primary interest in evaluating the effectiveness of instruction? In brief the chief of these would seem to be as follows:

(1) The extent to which students have acquired certain basic skills such as reading, writing, speaking, and calculating.

(2) The extent to which students have mastered some of the basic concepts and principles in each of the major areas of the curriculum.

(3) The extent to which students have acquired the attitudes and appreciations set up as objectives of instruction in particular courses.

(4) The extent to which students have mastered the techniques of study and modes of thinking in the major areas of the curriculum.

(5) The extent to which students apply creative imagination and critical thinking to real-life problems encountered in work, in citizenship, and in personal living.

(6) The extent to which students have gained a perspective on life and their relationship to it, with a conscious personal philosophy and set of values.

Changes in student behavior in these categories would seem to constitute the main test of the effectiveness of instruction in a particular college or university.

Measurement of student satisfaction. In addition to the evaluation of changes in basic skills, knowledge, attitudes, and appreciations, it is possible that the quality of instruction can be measured to some degree by the satisfaction or dissatisfaction expressed by the learners. Whereas student performance may be regarded as a cognitive outcome of instruction, student satisfaction may be looked upon as a conative or affective outcome. It could well be asked whether student satisfaction is really a valid dimension for the evaluation of instruction, on the ground that the changes in performance are the important things -- not how the student feels about it. The same could be said for merchandise which a consumer purchases in the market place. If a man's automobile actually transports him from place to place, why worry about his feelings in the matter? The good salesman knows that these feelings are very important; the good teacher is aware of this fact also. And in the case of the student, his attitude is important because it may well determine the extent to which he is motivated to continued learning.

The present state of affairs in measurement is such that we are far from sure that our measuring instruments are capable of detecting all of the many changes in behavior in which we are interested. This constitutes an additional reason for resorting to student satisfaction as a criterion for judging the quality of instruction, the assumption being that the student can make a subjective appraisal of educational changes which occur within himself even though these changes are not measurable by objective tests. The further assumption is made that if the student believes that he got a good deal out of a classroom experience he will express satisfaction with the experience; if he got little out of it, he will express some degree of dissatisfaction.

The difficulties encountered in measuring the outcomes of instruction in the classroom are duplicated in the field of counseling. At present we find it extremely difficult to measure the changes in behavior which result from counseling, but many investigators have reported that students felt that counseling was beneficial to them. Some, if not all, research workers have taken this expression of satisfaction as an indication that the counseling was effective. Much the same situation prevails in the field of evaluation of instruction. Perhaps the dimension of student satisfaction, since it is based upon a purely subjective appraisal, will some day be abandoned as a means of evaluating instructional effectiveness. For the present, however, it would seem to serve a useful purpose.

Measurement of supervisor satisfaction. Just as we find it necessary and desirable to use student satisfaction (based upon a subjective appraisal of classroom instruction) as a measure of instructional effectiveness, so it is possible to argue that certain aspects of the instructional process must be evaluated subjectively by a supervisor -- resulting in satisfaction if the instruction is good and dissatisfaction if the instruction is not good. As a general rule, one can probably assume that if a supervisor is dissatisfied with the instruction of a particular teacher, something will be done to change the quality of the teaching or the appointment will probably not be renewed. To be sure, sometimes little or nothing can be done about the work of an individual teacher; nevertheless, the satisfaction or dissatisfaction of the supervisor as expressed in comments or ratings may still be useful in the evaluation process. And while the satisfaction of the supervisor with the quality of a particular teacher's work may not be perfectly correlated with the quality of the students' learning, it is probable that the relationship would be positive. Here again, in view of the inadequacy of tests and other measuring devices, it seems necessary to regard supervisor satisfaction as one practical yardstick for determining instructional effectiveness.

Techniques of Evaluation

The evaluation of instructional effectiveness is usually concerned with the following types of questions: (1) How much have students learned in a stated period of time and in a particular course? (2) Does one method of instruction produce better results than another? (3) Is one set of instructional materials better than another? (4) Does instructor A produce better results than instructor B? (5) Is one type of course better than another in producing the desired educational results? (6) What effect do audio-visual aids have upon learning efficiency? (7) Is one type of course better in teaching intellectual content (ideas, principles, and skills) and another better in teaching attitudes and appreciations? (8) What is the effect of class size upon instructional effectiveness? (9) What is the relationship between student aptitudes and learning rate under different methods of teaching, e.g., lecture versus discussion? (10) What is the effect of student interest on learning efficiency?

Someone concerned primarily with the evaluation of particular facets of colleges and universities (as in this report) is not especially interested in the appraisal of the many different methods of teaching, content materials, media of communication (e.g., television), or adjustments which are made for individual differences among students. The important evaluation question is: Do students change in their behavior under the influence of this college, and can we measure this changed behavior in terms of skills, knowledge, attitudes, and appreciations? Nevertheless, many studies concerned with the ten questions listed in the first paragraph have employed techniques of measurement which would lend themselves to the evaluation of the instructional effectiveness of a particular college or university. And since the college's instructional impact is simply the total of effective devices employed by the members of the instructional staff, references will be made in the sections which follow to studies which were concerned only with the study of particular methods, types of students, or content materials -- provided the methods used would lend themselves to the broader use of evaluating the effectiveness of instruction on the college level.

Evaluation by opinions. It could be strongly argued that the only meaningful criterion of teaching effectiveness is change in student performance, and that, further, achievement examinations of various kinds should be the basic measures of changes in performance, supplemented, perhaps, with evaluations of those types of performance which cannot readily be so measured -- e.g., attitudes and appreciations. Since education is concerned with bringing about changes in the behavior of people, this line of argument would seem to have considerable validity. But, as pointed out earlier, the present status of educational measurement is such that we cannot rely exclusively on testing instruments which purport to measure changes in student behavior. We are forced to turn to a judge or observer as a means of evaluating materials, processes, and outcomes, even though judges may be less than perfectly reliable as measuring instruments. Until tests are developed which really measure the major outcomes of instruction, we must continue to rely on the evaluations made by an observer acting in the capacity of a judge. What techniques are available for making use of judges? What types of judges can be used in evaluating instruction, and what are the merits and demerits of the method?

One extensively used technique in this category of evaluation instruments is that of securing the opinions of students, colleagues, and supervisors, frequently by the use of rating scales, questionnaires, and check lists. Guthrie (25) has described the use of the rating method at the University of Washington -- the ratings being made by both colleagues and students. The ratings of teaching made by students are voluntary, but ratings by colleagues are required when a faculty member is recommended for promotion or if he is about to acquire tenure. The candidate for promotion or tenure status is asked to supply certain information about himself, to bring his bibliography up to date, and to make four nominations of colleagues to serve on his evaluating committee. The total committee usually consists of seven people, but does not meet as a committee. Each committee member, with the essential information before him, rates the candidate in question on 9 items, one of which is entitled "teaching effectiveness." The ratings are made on a man-to-man basis. The reliability for the scale was found to be .44. One of the interesting findings reported by Guthrie was that teaching effectiveness and research contribution had no significant correlation. The main features of the University of Washington system of rating teaching effectiveness are (1) colleagues do not necessarily observe directly the teaching in the classroom, (2) judgments are made upon the basis of a record prepared to describe the candidate's accomplishments, and (3) the ratings are of the man-to-man type.

Student ratings at the University of Washington are voluntary and are designed to be useful primarily to the individual instructors. One interesting finding in the University of Washington studies is that the correlation between the colleague rating of teaching effectiveness and student rating of the same item was found to be .48. Since the reliability of student ratings is .89 and that of faculty juries is .64 to .76, it is evident that students and faculty members do not have precisely the same idea as to what constitutes teaching effectiveness. It should be mentioned that the student rating scale of five items is entirely concerned with teaching effectiveness and also uses the man-to-man rating technique. Guthrie suggests that since students spend much more time observing the teaching effectiveness of a particular instructor than do colleagues, they are probably in a somewhat better position to evaluate this particular trait. Guthrie concludes that "there is no better criterion of teaching aptitude than teaching performance, and there are no better judges of teaching performance than students and colleagues" (25 , p. 115).

The position taken by Guthrie is to a considerable extent supported in the systematic analysis of student ratings undertaken at Brooklyn College and reported by Riley, Ryan, and Lifschitz (59). This study examined the theory and process of student evaluations and endeavored to discover what students are really looking for in their teachers as well as the degree to which the teacher fulfills this expectation. While recognizing limitations in student appraisals, the report emphasizes that teaching is a two-way process and that the judgment of students has considerable validity and reliability in measuring the effectiveness of the process.

The strengths and weaknesses of the opinion method of evaluation of teaching effectiveness are very much the same as they are in the areas of curriculum and objectives evaluation. The measuring instrument is the human being, serving as observer, recorder, and judge. The method can be only as good, or as bad, as is the human being functioning as a measuring instrument.

Systematic observation of teaching materials and processes. The opinions of students, peers, and supervisors, usually obtained by means of rating devices, are in many (if not most) cases based upon general impressions, hearsay, and second-hand sources of information. In other words, the opinion method of evaluation may be deficient not because of shortcomings in the method itself, but rather as a result of the manner in which opinions are collected. It has been argued that if ratings were made on the basis of direct observations of behavior, they would prove to be more satisfactory as criterion measures and as predictors of later behavior.

Because of the dissatisfaction with the generalized information on which many ratings or opinions are based, it has been suggested that there should be more direct, systematic observation of the teaching process, of the materials and techniques used in teaching, and of the responses of students while they are present in the classroom. The use of direct observation involving a time sampling procedure is illustrated in a study by Johnson (33). Two observers who had had a considerable amount of pre-experimental practice together rated 13 teachers every five minutes of a twenty-minute session. Six such sessions were observed during a summer session, thus yielding 24 ratings by each observer for each teacher. The results of the five-minute ratings were added and a mean determined. The final grade was the sum of the averages of effect of teacher on class, reaction of class to teacher, adequacy of student, and outcome of general experience. (These categories and how to rate them are fully described in the article.) The reliability of scoring by use of the scale was .96. The validity of the method is not subject to check, but the results do indicate high agreement between raters. While the study does make use of ratings, it is significant that these ratings are based upon direct, systematic observations.

Teacher-made examinations. The construction of examinations is a part of the work of every teacher. While these examinations are constructed and administered primarily in order to grade the work of the student or to motivate and aid his learning, it is not uncommon for teachers and administrators to make evaluative comments about particular methods, content materials, or types of courses on the basis of the performance of students on these teacher-made examinations. Sometimes these examinations cover only a particular unit of a course; at other times they may cover an entire semester or year of work. And if a teacher moves from one institution to another he may make comparisons between the students in College A and those in College B on the basis of comparative test scores in his courses. How valid is the teacher-made examination as a means of evaluating student performance in a particular institution or in different institutions?

Much has been written about the strengths and weaknesses of teacher-made examinations as evaluation devices. Frequently these tests are constructed on short notice, with no try-outs of individual items, with brief and often inadequate directions, and with incomplete coverage or poor sampling of the subject-matter taught. On the other hand, a teacher may spend a great deal of time in preparing examinations, in accumulating good items, and in analyzing the test results. In brief, the quality of a teacher-made examination is very often unknown; hence any results obtained by its use must be interpreted with great caution.

Regardless of the quality of the teacher-made examination, all these examinations lack one very important thing, namely, norms of performance. The short quizzes, hour examinations, and final examinations in chemistry at a particular institution may be of very high quality, but its students' performance in chemistry cannot be compared with students' performance at another institution, even though both may use the same grading system. One might say, therefore, that teacher-made examinations are useful primarily within a particular institution.

A very interesting use of teacher-made examinations in the evaluation of instruction at particular institutions is described by Haggerty (48 , pp. 44-50). A group of mathematicians was asked to rank the quality of the mathematics instruction in 57 colleges and universities, using the final examination papers in freshman and sophomore mathematics courses as the raw data for making this evaluation. The committee developed for each course a content pattern and a scale of student mastery by which they rated the examination papers of individual institutions. After studying the character of the examinations themselves and the performance of students in these examinations, and knowing the background of each mathematics instructor, they developed a quality-of-mathematics-instruction score for each institution. It was the judgment of the committee of mathematicians that this procedure provided a better means of evaluating the quality of mathematics instruction than would be the administration of one uniform examination for each course in all the institutions included in the study. The results of the mathematics evaluation study were ultimately included in the over-all measure (Criterion VIII) used to evaluate the institutions included in the North Central Study. It would be of considerable interest to have this type of study duplicated in other courses. In the final analysis the evaluation achieved in the mathematics study was, of course, a rating, and is, therefore, subject to the shortcomings of the rating method. A comparison of an evaluation score achieved by a committee (as in the mathematics study) and that obtained by a set of standardized achievement tests for the courses investigated would be of considerable interest and value.

Standardized achievement examinations. The standardized achievement test has been used very widely as a measure of instructional effectiveness. In some cases standardized tests have been used in well-designed studies; in others, the results of testing programs have been analyzed and conclusions about teaching effectiveness drawn when the original purpose of the testing program was to measure pupil performance without particular reference to instructional effectiveness. Needless to say, many of the conclusions about teaching effectiveness based upon test scores are lacking in validity. It is precisely because of the temptation to draw unwarranted conclusions about teaching effectiveness from the results of testing programs that many teachers view with misgivings the introduction of large-scale testing programs. Nevertheless, the standardized test does place the emphasis on student performance and hence has been widely used as one means of evaluating instruction.

The peculiar advantages of the standardized test for measuring instructional effectiveness are very much the same as the advantages generally listed for using standardized tests in any situation -- the availability of norms, known test quality as indicated by indices of validity and reliability, ease of administration and scoring, and comparability of results with those obtained in other situations. This is not to say that all these advantages have necessarily been realized in every study of teaching effectiveness. As a matter of fact, the availability of certain tests has probably meant that they have been used in situations where locally constructed tests would have better served the purpose.

One of the chief limitations of the standardized test as a measure of instructional effectiveness is its possible lack of validity in the particular situation in which it is used. A test may purport to measure outcomes of instruction in college chemistry and may represent what a group of experts believe should be tested. If, however, the test emphasizes primarily facts, problems, and laboratory procedures, whereas the course has emphasized primarily the understanding of basic concepts in chemistry and of the scientific method, the test will not be valid for the purpose. Students in the "facts and problems" course may appear to have been taught better than students in the other course only because the test was more appropriate for them. In short, the standardized test must be valid for the course, or courses, in question if it is to be used as a measure of instructional effectiveness. The mere fact that it has a manual of directions, tables of norms, and indices of reliability does not make it always appropriate for testing educational outcomes.

Indeed, the norms may actually have little value in a local study. The chief merit of using the standardized test in an experimental situation may be that the items are likely to be of superior quality and in such form that the test can be readily administered. In brief, the usefulness of the standardized achievement test as a measure of instructional effectiveness must be judged in terms of the demands of a given situation.[*]

The pre-test and post-test method. Teacher-made and standardized achievement tests are most frequently used to measure instructional effectiveness by determining how much students gain in test scores during a period of instruction. This procedure was used extensively in the study of general education reported on by Dressel and Mayhew (14 , pp. 32-174). The typical procedure in the General Education study was to give the achievement test at the beginning of the course and to readminister the test at the conclusion of the term. The average gain in test scores was then taken as a measure of what students learned in the general education courses investigated in the study. An illustration of the results obtained and the methods used is presented in the discussion of the test of critical thinking in social science (14 , p. 59). "Gain scores" are presented for eleven different colleges and for the total group, with a breakdown for students of five different levels of initial achievement.

The gain-score results for students of different levels of initial achievement underscore the importance of considering initial achievement when comparing the performance of students in different classes, in different courses, and in different colleges. If only post-test scores were used in a particular situation it could well happen that differences in student performance would be attributed to quality of instruction when actually the initial standing of the students was the factor accounting for the differences. Allowance for differences in initial ability can be made, if appropriate scores are available, by covariance analysis.

---

[*] Since standardized achievement tests have been used in so many studies, it is not deemed necessary to cite specific instances illustrating their use.

Another very interesting and important fact brought out in the Dressel and Mayhew report is that, in general, students in the lowest quartile of initial achievement appear to show the greatest gain and those in the top quartile of initial achievement appear to show the least gain. The fact that many courses seem to cater to low-ranking students, the familiar regression effect, the lack of a high ceiling for the test, and the nature of the test items may account for this phenomenon. Whatever the cause, it is obvious that the initial level of student achievement must be known if one is to draw valid conclusions about the effectiveness of instruction in different classes, courses, or institutions.

A study which illustrates the experimental evaluation of both content and methods of instruction, using the pre- and post-test methods, is that of Bond (8). The purpose of the study was to determine "changes effected in certain college students during presentation of a teaching unit with objectives consistent with those of general education," in comparison with changes in students who studied a block of subject matter in genetics organized and approached from a more conventional point of view and emphasizing more the physical basis of heredity. Pre-tests and post-tests were given in the following areas: knowledge of concepts, reduction in acceptance of common superstitions, ability to apply principles, critical appraisal of authorities, accuracy of interpretation of data, and changes in opinions about certain international questions and certain nationalities. Differences between the experimental and control groups were controlled statistically, and both groups had the same teacher. The results were evaluated on the basis of changes from initial to final testing. The chief finding of the study was that the experimental group appeared to gain more in the area of general education without any loss in acquisition of facts and understanding of concepts when compared with the control groups.

Attitudes and appreciations tests and scales. One fact upon which all educators and evaluation experts agree is that the measurement of outcomes of instruction should not be limited to the testing of facts, skills, and intellectual concepts and ideas; a complete measure of teaching effectiveness should include changes in attitudes and appreciations as well as in the intellectual sphere. This is true for the teaching in the natural and social sciences as well as in the humanities. Scientists are very much concerned that students should develop a scientific attitude toward the solution of problems and grow in the power of critical thinking (which is not purely intellectual). Social scientists are interested that students develop a sensitivity to social problems, that they approach these problems with open-mindedness, and that they develop such traits as tolerance, willingness to understand the viewpoints of others, and willingness to compromise on controversial issues. Humanists want students to understand and appreciate the role of values, to develop a sense of the beautiful, the creative, and the spiritual in human life. Some of these outcomes are associated with the knowing of facts and ideas and hence are indirectly measured by means of tests which include only intellectual content. However, there is a considerable list of outcomes in the area of attitudes, appreciations, and modes of thinking which are not measured by tests of the cognitive type. What techniques are available for the measurement of these important outcomes of instruction, and how successful have these measurement techniques proved to be?

Efforts to measure changes in attitudes, appreciations, and modes of thinking have been included in several testing programs during the past twenty-five or thirty years. In addition, a good deal of research and of test or scale construction has been done by individual institutions and test construction agencies. One of the most recent testing ventures in this field on the college level has been the study of evaluation in general education reported by Dressel and Mayhew (14, pp. 174-241). In this study an attempt was made to measure, in students' critical thinking and attitudes toward ideas and institutions, social groups, individuals and self, changes occurring during the time that students were enrolled in general education courses. Strictly speaking, this project was more concerned with curricular problems (general education) than with teaching techniques, but the instruments developed could readily be used in research on teaching effectiveness. Indeed, it should probably be said that in this study of general education the investigators were really measuring teaching effectiveness as influenced by a combination of curricular content and teaching skill.

One of the instruments developed was the Test of Critical Thinking, designed to measure the ability to define problems, to select pertinent information, to recognize assumptions, to formulate and select relevant hypotheses, to draw conclusions validly, and to judge the validity of inferences. Extensive studies were made of the test, including correlations with well-known intelligence, reading, and achievement tests; with ratings of critical thinking ability by self, peers, counselors, and teachers; with grade-point averages; and with other tests purporting to measure thinking ability. On the whole the test seemed to perform about as would be expected for such an instrument.

The main use made of the Test of Critical Thinking in the General Education study was to measure change in critical thinking scores during enrollment in various types of courses, particularly courses in general education. The principal results are those reported for changes occuring during the freshman year for students enrolled in seven colleges. All groups gained, but the greatest gains were made by initially low-scoring groups and the smallest gains were made by the initially high-scoring groups.* The investigators did not find any definite relationship between type of course taken and student performance in the Test of Critical Thinking. Nevertheless, the type of educational outcome measured by the Test of Critical Thinking is believed to be significant, and hence would appear to represent a field which should be investigated much more intensively.

A second instrument of a new type developed in the General Education study is the Inventory of Beliefs; it consists of 120 statements with which the student can (1) strongly agree, (2) agree, (3) disagree, or (4) strongly disagree. The committee which constructed the inventory prepared items with which (in their judgment) students should disagree. Many studies were made of the Inventory of Beliefs, including correlations with intelligence and achievement tests and with other tests purporting to measure reasoning and critical thinking. The investigators concentrated their attention on the mean gains of students in 13 colleges during one year of college experience. Again, there were decided differences from college to college, and the greatest gains were made by the initially low-scoring group. The results obtained in the General Education Study are not impressive, but they do suggest strongly the need for further research.

A third instrument developed is the Human Relations Test; it is designed to measure the students' attitudes toward persons and processes, and uses the situational type of objective test question (60 in all) as the testing medium. The attitudes which could be exhibited were classified into five major categories as follows: (1) hard-boiled autocrat, (2) benevolent autocrat, (3) laissez faire, (4) democratic, (5) resorts to "Experts" or authority figures. The scoring of the test is based upon the extent to which the respondent exhibits the democratic attitude. The test was found to have fair reliability (.64 to .73) and appears to possess considerable face validity. Again, the investigators were concerned with gains made in the test scores during one year of college experience. All the mean gains for nine colleges were very small. Once more, while the results in this particular study may be disappointing, it would appear that the measurement problem studied is worthy of much additional time and effort.

---

*The reasons for this are discussed on pp. 112 of this report.

Another type of approach toward the measurement of attitudes and appreciations is represented in the Humanities Participation Inventory developed in the General Education study. The basic assumption in this type of test is that attendance at concerts, operas and plays, the reading of books, the playing of musical instruments, and many similar activities are indications of how much people appreciate these activities. The testing technique is to ask how frequently the respondent has participated in an activity, how much the activity was enjoyed, and how seriously it was undertaken (three degrees in each of the three categories). The results obtained by the use of the inventory were similar to those obtained with the instruments described above; small gains during one year of college experience, higher gains by low-scoring than by high-scoring groups, and no definite relationship between type of college experience and scores on the inventory.

Taken as a group, one would have to say that the efforts to measure attitudes, appreciations, and modes of thinking have not been especially successful when judged by the changes which occur in students' scores during a year of college experience -- assuming, of course, that real changes in behavior did occur as a result of college training. There are a few exceptions, such, for example, as the study reported by Timmons (69). The measurement problems in this area are formidable, but the stakes are high; only the initial beginnings have been made in an area of measurement which holds tremendous significance for education. There is much room for bold, new thinking, for creative effort and imagination. The results obtained to date may be disappointing, but the challenge to continue the research is very great. Viewed over a period of time, it is very obvious that educational forces do produce changes in attitudes and appreciations. The problem is to develop instruments sensitive enough to detect changes which occur in relatively short periods, such as a year. When this is accomplished, a real contribution will have been made toward the evaluation of teaching effectiveness.

Follow-up studies. The follow-up study has been suggested as a means of evaluating a number of facets of college work. It would seem that it should also be applicable to instructional evaluation. The possibility exists that the changes in behavior produced in students are not discernible or detectable during the time the course is being taken or at its conclusion. And if the satisfaction of the student is used as a dimension of evaluation, it is possible, of course, that students may be in a better position to evaluate the effectiveness of an instructor after an interval of time has elapsed, rather than during the course itself. It would therefore seem that follow-up studies of students who have taken particular courses may be helpful in evaluating the effectiveness of instruction.

Follow-up studies may be of two general types. First, there are studies which actually describe what students enrolled in particular courses do or fail to do: How many students taking an introductory course with Professor X major in the field, compared with the number majoring in the field who studied with Professor Y? Second, there are studies which solicit the opinions of students, alumni, and former students about the quality of the teaching or the value of the content of courses taken some months or years earlier. Both types of studies have possibilities and limitations as techniques for evaluating the quality of instruction.

The type of study which follows the actual behavior of students is illustrated in a report by Beaumont (2). This investigation attempts to measure the effectiveness of instruction in elementary psychology by determining the quantity and quality of advanced work done by students prepared under six different instructors of elementary psychology. After giving some consideration to the factor of differences in intelligence of the students and to the fact that some were required to take advanced courses, it still seemed that some instructors were more successful than others in encouraging and preparing students to take advanced courses. Beaumont suggests the method, that if used judiciously, may provide a means of judging the effectiveness of teaching which is more objective and factual than more conventional methods.

The difficulties inherent in the above type of study are fairly obvious. Because students are not entirely free to choose their courses, and because of individual differences in aptitudes, interests, motivation and personal attributes, the results of a follow-up study which describes or tabulates what students actually do must be interpreted with great caution. Certain factors such as differences in intellectual aptitude can be handled statistically, but some of the other factors would be very difficult to control. Nevertheless, teachers of advanced subjects do frequently say that the quality of earlier teaching is discernable in the work of their advanced students. If proper controls are exercised, it would seem that this type of follow-up study would be worthy of further investigation.

A more comprehensive type investigation is illustrated in the publications of Knapp and Goodrich (39) and Knapp and Greenbaum (40). These two reports are concerned with the success of two large groups of students; conceivably, quality of teaching could be one factor influencing the results, although it must be said that the studies provide no actual data on this point. Since the techniques used in these two studies are described fully in Chapter VIII, space will not be devoted to a description of them here.

Two important factors which should be considered in interpreting the findings in these studies are: (1) social class of the homes from which these outstanding students come (primarily middle class), and (2) the difference in emphasis between colleges and universities in their programs of instruction, universities being much more professional. A part of the difference in the productivity indices of various institutions can be accounted for by these factors rather than by differences in quality of instruction. Further comments on these two studies are presented in Chapter VIII.

The limitations inherent in the follow-up study as a measure of instructional effectiveness have been pointed out very clearly by Haggerty (48 , pp. 29-36). The University of Minnesota, with the help of a grant from the Carnegie Foundation for the Advancement of Teaching, made a study of 26,000 graduate students enrolled in seventeen universities, twelve of which were in the North Central region. These students came from approximately nine hundred undergraduate colleges and entered graduate work between September 1926 and June 1931. The main criteria for judging the graduate college success of these students were the number of graduate hours completed satisfactorily, the number of hours failed, the record in graduate examination, and the degrees earned. The combined records in graduate work of the students from each undergraduate college were then determined and tabulated. As one would expect, there are significant differences among the groups of students in terms of their success in graduate work.

As pointed out by Haggerty, the factors which must be considered in interpreting the results are: (1) the percentage of graduates of a particular college going to graduate school, (2) the fact that only those graduates taking graduate work in the seventeen universities included in the project were studied, (3) the variation in the grading standards of the several graduate colleges, and (4) the fact that the percentage of graduate work completed is quite high (86 to 98 percent), thus reducing the variability in the criterion measure. Since there did not seem to be opportunity to correct for these factors, no attempt was made to include data from the follow-up study of graduate students as a part of the North Central Association investigation of accrediting procedures. One rather disquieting finding (possibly a direct result of the inadequate design of the study) was that students from non-accredited colleges seemed to do about as well in graduate school as did those from schools belonging to the North Central Association.

Ideally, if a follow-up study of graduates going on to professional or graduate study is to be successful it should account for all possible members of the experimental population (or a true random sample), all should attend the same or very similar graduate or professional schools, and all should be judged by the same standards of performance. Since these conditions are not easy to meet, it can be seen that, despite its attractiveness at first appearance, the follow-up study is subject to serious limitations as a means of evaluating teaching effectiveness.

Evaluation Needs

The problem of evaluating instructional effectiveness has been studied over a number of years in nearly all aspects of the educational program and at every level of instruction. A very good summary of the studies concerned with the evaluation of teacher effectiveness and the relationship between various teacher characteristics and teaching successes has been prepared by Morsh and Wilder (46). There would be little point to repeating in this report a summary of the studies of teaching effectiveness which have appeared in the literature. Morsh and Wilder have done this job quite thoroughly, including the significant work on the elementary and secondary as well as the college level. The important questions are "Where do we go from here? What is needed most at this time?"

Better measures of achievement built around particular instructional objectives. A great deal has been accomplished during the past thirty-five or forty years in the construction of educational achievement tests. Tests are available for the measurement of such skills as reading, writing, and calculating, and for the measurement of achievement in most subject-matter areas of the curriculum, at least through the junior college level. Many of these examinations are of superb quality, well standardized, and easily administered and interpreted. Several of them, like the Iowa Tests of General Educational Development, are available in several forms, thus making possible the measurement of growth over a period of years. Granted that much has been accomplished in educational achievement testing, what is still lacking? Why are present tests not universally accepted as valid measures of teaching effectiveness?

It would appear that one reason for not accepting presently available achievement tests as measures of teaching effectiveness is that they do not appear to measure student behaviors closely enough related to the particular objectives sought in specific courses or educational programs. The fact that the American Council on Education sponsored an extensive study of general education which called for the construction of testing instruments in such areas as the natural sciences, the social sciences, and the humanities indicates that there was a feeling that no satisfactory instruments were available for measuring what responsible experts, or even particular faculties, considered important outcomes of general education. To be sure, there are tests of knowledge in science, social science, and the humanities, but for the most part these tests seem to strike an average of what a large number of people want to have measured in a particular field. On the college level, particularly, there is a good deal of individuality in the teaching of courses. Many colleges, for example, give a core course in social science; yet when the instructors examine a test purporting to measure knowledge and understanding of social science, it doesn't quite suit their purpose; hence they proceed to build tests of their own. These instructors would view with considerable misgiving any attempt to measure outcomes of instruction in their course, if the test emphasized educational objectives different from those which they had in mind. Briefly, in many cases tests are available which have the right name, but not the right emphasis!

It is fairly obvious that if instructional effectiveness is to be measured by means of state or national testing programs, there will have to be greater standardization of curricular and course objectives. In some areas of the curriculum, standardization may be neither feasible nor desirable. Perhaps a beginning could be made in such fields as mathematics, physical science, biological science, history, and certain of the foreign languages. If, for example, a committee of mathematicians could agree that certain skills, concepts, and problem-solving abilities should be learned during a year of study of college mathematics; and if a test, or tests, were constructed to measure these outcomes, a beginning would be made toward the building of an instrument which could be used to measure instructional effectiveness in different types of courses, taught in different colleges and under different instructors. To build such instruments in certain social science and humanities courses would be out of the question for the present because of the uniqueness of objectives stressed in many of these courses. Since instruction is more standardized in mathematics and the sciences than in other branches of the curriculum, it would seem that the first attempts to measure known and agreed upon objectives should be in these subjects; at least it would be worthy of exploration.

If it should not be feasible to construct one test of college mathematics or one test for each of the standard science fields, perhaps it would be feasible to build tests yielding part-scores of a wide variety.

Each institution could then determine which parts were suitable for use in its particular situation. Norms could be provided for various combinations of part-scores, thus making available a standardized instrument which could be used in the evaluation of instruction. This, too, is an approach which would be worthy of exploration.

In those areas where subject-matter is relatively unstandardized, the chief hope apparently lies in building excellent locally constructed examinations, probably with the help of testing experts. If instructors and testing experts could sit down together to identify objectives and then build test items to measure them, progress could be made toward the goal of securing instruments acceptable as valid measures of teaching effectiveness. At present, Educational Testing Service is developing portfolios of test items obtainable by individual teachers for use in their own examinations. If enough common items were used by instructors in a number of different institutions it is possible that some basis of comparison could be worked out. It is not certain that this would really solve the problem of providing comparable measurements for various institutions, but it might well constitute an important intermediate stage in the development of comparable measurement programs.

Still another alternative is to measure the intellectual processes necessary for effective work in a given field -- the ability to understand what is read, the ability to make inferences from the content, and the ability to estimate the consequences of alternative solutions to problems. These are, of course, among the abilities measured by the General Educational Development Tests. While these tests are good as far as they go, many faculty members feel that on the college level they tend to measure outcomes so general as to be inadequate for measuring instructional effectiveness.

Despite the progress which has been made in achievement testing, few college administrators would, for example, use the test scores of students to appraise the teaching effectiveness of different instructors teaching sections of the same course, for in many cases the testing instruments are not adequate measures of the stated objectives of instruction. The building of standardized tests in some areas (e.g., mathematics and general education outcomes) and the improvement of locally constructed examinations in all other areas would seem to be essential if progress is to be made in using achievement tests as measures of instructional effectiveness.

Better test norms. Hand in hand with the need for testing instruments which really measure the outcomes of instruction emphasized in particular courses and institutions is the need for adequate norms for these tests. Assuming that an adequate test is available for measuring the outcomes of instruction in first-year mathematics, satisfactory norms are necessary if the test is to be useful in measuring instructional effectiveness, particularly in small institutions. In a college with 800 students and 20 instructors in college algebra, the course supervisor may be able to use a good test to measure the instructional effectiveness of these 20 teachers even though no norms are furnished by the test builder. The small college, with 40 students of algebra under one teacher, is faced with a different situation. The supervisor here needs to know the average increase in students' knowledge of college algebra during a year of instruction in other colleges of his type. Furthermore, he should have these gain scores for students of varying levels of ability. Few tests at the present time have norms for colleges of various types, broken down within colleges according to the ability level of students. Unless the teacher in the small college has this sort of information he can hardly be expected to be enthusiastic about using achievement tests as measures of his instructional effectiveness. A beginning along this line has been made in achievement tests for various elementary school subjects such as reading, but very little progress has been made in providing this type of test norm on the college level.

Better techniques for assessing changes in attitudes and appreciations. There have been a number of studies of changes in student attitudes and appreciations, and a variety of techniques for measuring these outcomes of instruction have been developed, including attitude scales, inventories of interests and beliefs, tabulations of activities of various kinds, ratings, and measures of changes in knowledge of facts and concepts. Despite the interest in this field of measurement, the number of studies actually completed on the college level is not large. Compared with the time, effort, and resources which have gone into the achievement testing field, little has been done to measure changes in attitudes and appreciations resulting from the typical college course.

Granted the inadequacies of the techniques which have been developed to date, it would, nevertheless, be desirable to continue investigations like those of the General Education Study of the American Council on Education. Unless time and effort are spent in thinking about a problem, it is not going to be solved. If attitude scales, inventories of beliefs, and participation inventories were adapted for use in a wide variety of courses and were used in a number of different colleges, it is likely that some progress would be made in improving the instruments and in focusing attention on the importance of thinking about changes in attitudes and appreciations as outcomes of instruction. Most college teachers believe that their courses do serve to bring about changes in points of view, in values and appreciations, but that measurement of these changes is so complex that they are not subject to evaluation. The wider use of instruments of the sort developed in the General Education Study might lead to a better understanding of the problem and its eventual solution.

This area should challenge the imagination. It is closely related to the field of personality assessment, which, it is generally agreed, is far less advanced than intelligence and achievement testing.

It would be unrealistic, therefore, to minimize the inherent difficulties involved in developing better measures of attitudes and appreciations. Time and space do not permit a full elaboration of the subject, but it would seem that a combination of the cognitive and affective approaches might prove to be more successful than the techniques used thus far. For example, it has been found that an informational type of test can be used as an indirect measure of interests. The things that people know about and comprehend to some degree reflect their interests. Perhaps there are also things which one knows about or understands only or primarily because he has certain attitudes and appreciations. There are, no doubt, many other ideas and approaches which should be tried. Knowing the history of educational and psychological measurement, one can safely predict that the problem of the measurement of attitudes and appreciations will eventually be solved. But until it is, it will not be feasible to rely very heavily on measures of attitudes and appreciations as means of evaluating instructional effectiveness.

Better measures of critical thinking and problem-solving. One often heard justification for liberal education is that it teaches students to think more critically and to use appropriate methods in attacking new problems. The methods of gaining new knowledge in the humanities, the social sciences, and the natural sciences are different. Liberal arts educators believe that the student should study subjects in each of these areas in order, partly at least, that he may gain an understanding and appreciation of these different methods. It is further assumed that the student will later be confronted with problems in which these different methods can and should be applied. Many, if not most, college teachers would say that their courses do teach these skills and abilities, but objective evidence on the question is not very plentiful. It would be very difficult at this time to apply measures of critical thinking or of problem-solving methods to the teaching situation and to say on the basis of results obtained that Professor A is more successful than Professor B, that course M is better than course N, or that College X is better than College Y. In other words, we are pretty confident that college teaching does result in changes in critical thinking and related skills, but we have no sure-fire method of measurement to apply. Just as is the case with the measurement of attitudes and appreciations, a number of studies have been made, several different techniques have been tried, but the problem remains. Since the measurement of these educational outcomes is of the greatest importance to higher education, it is essential that research on it continue and that it be given wide attention and support. Unless it is pursued relentlessly, the problem will not be solved.

Better methods of observation and rating. The present status of evaluation of teaching effectiveness is such that direct observation of the teaching process by students, colleagues, supervisors and teams of experts is essential in a total program of instructional evaluation, and perhaps this will continue to be the case for some time to come. One may wish for tests which would measure the changes in student behavior brought about in classrooms and laboratories, but at present such tests are unsatisfactory. The discussion of the strengths and weaknesses -- chiefly the latter -- of tests, inventories, and rating scales as measures of teaching effectiveness makes it clear that efforts to improve the techniques of direct observation should not be slackened or abandoned. On the contrary, every possible effort should be made to improve them.

There probably are definite limitations to the degree of improvement which can be brought about in the observations and ratings made by students and colleagues. Students are limited in their experience and, furthermore, are not in a position to evaluate all phases of teaching effectiveness. They can report their feelings about teaching but may or may not be very good judges of how much they really learn under different teachers. Colleagues are seriously limited in the judgments which they can make because of the scarcity of their opportunities for observing the teaching of their colleagues. One may deplore this situation, but it does exist and will probably continue to exist in the future.

The chief hope would seem to lie in the more efficient use of supervisors and teams of experts in making observations and ratings of teachers in action. The availability of motion pictures, audio-recording devices, and television should make it easier in the future to use observation and ratings as techniques for the evaluation of teaching. Direct observation of specific aspects of the teaching process by trained observers who take their job very seriously, who have defined precisely what they are looking for, and who have developed a rating scale for each facet of teaching to be observed would seem to possess real promise. One major difficulty with rating procedures in the past has been the fact that ratings are made on the basis of information which is too little and too fragmentary. An organized program of evaluating the teaching in all sections of a course such as freshman English by having trained observers rate important aspects of the teaching process either by being present in the classroom or by listening to and looking at recordings -- e.g., TV kinescopes -- of actual teaching, would constitute a real step forward in the evaluation of teaching. Teachers are frequently appointed, promoted, or dismissed on evidence which is largely second-hand in nature. It would be refreshing to see a program of instructional evaluation based exclusively on first-hand information.

While student ratings have proceeded in some institutions to a point of considerable refinement, most of them need a new emphasis. Too many appraise the instructor rather than the instructional process. The good teacher is not merely an actor whose performance is to be observed and criticized. Rather, he and his students are partners in a common intellectual quest, and both partners must be judged, particularly on effectiveness of their shared activity. In such an evaluation situation students are assessing themselves as well as the teacher and in the end expressing a deeper judgment about the effectiveness of the learning experience. Some colleges are now in the process of revising their evaluation instruments to include this new emphasis.

Better definitions of good teaching. When all is said and done about the importance of better techniques and interpretive data in the evaluation of teaching, the fact still remains that the greatest need is a more precise definition of what is meant by good teaching in each subject. In a subject such as reading, the important objective is to get pupils to read better; in freshman English it may be to get students to write better; but what about a course in American government, or chemistry, or psychology or art? Does each teacher of these subjects know what is meant by good teaching in each of these fields? Should the student learn facts, solve problems, develop the scientific attitude, enjoy a new intellectual experience, learn to appreciate something which has not been appreciated before? Part of the confusion in the field of instructional evaluation is that we aren't exactly sure what we want to measure. We succeed in arriving at some average judgment as to what we mean by good teaching and then proceed to rate or test the type of performance indicated by this average judgment. But is this average judgment a good judgment? In many cases we don't really know. What we ought to have for each course of the curriculum is a clear, positive statement as to what is meant by good teaching in that subject (cf. Chapter I on Objectives). This should be very helpful in the evaluation of instruction -- and one can be fairly certain that it would have a wholesome effect on the quality of teaching, something which is, after all, the real objective of much, if not all, of what we are trying to do in higher education.

One possible approach to the problem of specifying what is meant by good teaching is through the application of Tucker's "schools of thought" technique, described in the chapter on objectives. In this kind of a situation, classroom materials, brief descriptions of classroom activity, short film or sound recordings of classroom activities, etc., could be used to construct items which might be judged by experts in the field as indicative of the quality of teaching. The analysis would result in clear identification of the different views or schools of thought as to what constitutes good teaching. A value judgment might then have to be made as to which view to accept, or possibly separate measures would be developed to represent the different schools of thought.

Summary of Evaluation Needs

The techniques and interpretive data needed for the evaluation of teaching effectiveness on the college level are summarized in the accompanying table. Perhaps these needs can best be illustrated by describing what a course supervisor of social science would like to have for measuring the teaching effectiveness of ten teachers, each in charge of three sections of thirty students per section.

First, he would like to have a test of social science achievement which he and other competent authorities agreed measured the facts, concepts, attitudes, appreciations, and intellectual processes which students might reasonably be expected to acquire as a result of studying social science on the college level for one year. The test should, of course, have range enough to measure the lowest degree of social science achievement for the weakest student and the highest degree of achievement for the best student at the end of the course, and it should have part scores for the different types of educational outcomes. He would want to give one form of the test at the beginning of the course and another form at the end, using the method of covariance to correct for differences in the proficiency of students in the thirty sections at the beginning of the course. The gain scores (relative to initial ability) of the students would then constitute his chief measure of the teaching effectiveness of the ten instructors in the course. If available, he would, naturally, like to compare the gain scores of his students with those of comparable students at other institutions enrolled in the same type of course.

Second, recognizing that even the best of tests have limitations, he would want to have direct observations made at various times of the teaching of each of the ten instructors in the course. This he would do by having a team of three experts rate each teacher once a week for three five-minute periods during that particular hour. To avoid disturbing the class discussion, he would make his observations by having each class televised (closed circuit) and having the proceedings from each classroom channeled to one central viewing room. The observers would be trained to watch for specific things, e.g., how many students were paying attention, how many asked questions, what types of questions were asked, whether there was discussion among the students themselves, the atmosphere of the class (teacher-dominated or student-centered). With a previously established system of weights, each of these attributes of teaching would be scored, both subscores and total scores being reported. The results of this direct observation would then be carefully correlated with the gain scores obtained at the end of the course in the social science achievement test.

Third, he would ask the students in each section to rate their teachers on the five to ten aspects of teaching which, in his judgment and that of other experts, students are in a position to evaluate. He would carefully explain to the students the nature of the project, the meaning of each rating dimension, and the procedures to be used in making a mark on the rating scale, possibly preceding the explanation by helping them construct a man-to-man scale to be used in carrying out the rating task. A composite rating would then be computed for each teacher and correlated with the achievement test scores and the direct observations.

Finally, he would want to make a follow-up study of the students in each of the sections, noting how many took additional courses in social science (assuming randomization of interests in each of the original sections), how many majored in a social science department, how many went on to graduate work and how they performed in advanced tests such as the Graduate Record Examination. Likewise he would want to assess the activities of alumni who, though perhaps they took no advanced courses, should show in their daily behavior refined tastes and insights growing out of the course. The results of the follow-up study would be interpreted with extreme caution and would be used primarily as supplementary data in the evaluation of the work of each teacher.

Adequate examinations, careful observations by experts, carefully gathered student opinions, and thorough follow-up data would contribute immensely to the evaluation of teaching effectiveness in classrooms and laboratories. Very few colleges have ever carried out the type of project described here, even on an experimental basis -- certainly not on a routine, practical basis! One cannot but feel that if it were to be done it would have an effect on the quality of college teaching which would be without precedent in the history of higher education in America.

Summary Table for Chapter V

Evaluation of Instructional Effectiveness

Evaluation Problems →	Measurement of Student Performance	Measurement of Student Satisfaction	Measurement of Supervisor Satisfaction
Available or Existing Measures and Techniques for Evaluation	1. Teacher-made tests of knowledge 2. Standardized tests of knowledge 3. Control group experiments using pre- and post-testing 4. Follow-up* 5. Records of behavior outside classes (e.g., number of library books used) 6. Ratings of student performance	1. Questionnaires 2. Rating scales 3. Attrition rate	1. Questionnaires 2. Rating scales 3. Systematic observation of teaching materials 4. Systematic observation of teaching process
Measurement Needs	1. New and improved tests -- especially of attitudes and appreciations 2. Improved design of control group studies	1. Objective indices of satisfaction	1. Improved instruments for recording observations of teaching process
Available Interpretive Data	1. Normative data on published tests		
Needed Interpretive Data	1. More extensive normative data on existing tests 2. Normative data for all other measures 3. Normative data on performance in relation to ability	1. Normative data for all measures	1. Normative data for all measures 2. Information on relationship between specific teaching materials and studen gain 3. Information on relationship between specific teaching procedures and student gain

---

*See Summary Table for Chapter VIII, column 1.

PRECEDING PAGE BLANK- NOT FILMED

Chapter VI

THE STUDENT BODY

Introductory Statement

Even the most research-minded faculty member would hardly deny that the student body represents an important part of the university. In spite of this, there appears to be little evidence in the literature to indicate that characteristics of the student body have been used directly to evaluate an institution. Among institutions having essentially the same aims and offering similar programs, such things as the level and type of students attracted do represent important and unique items of information in an evaluation of the kind of job the school is doing. Thus, if cognizance is taken of appropriate contingencies, characteristics of a student body can provide valuable direct evidence of the quality of a college or university.

A study of the characteristics of a student body is also important to the evaluation of an institution in other ways. For example, in the evaluation of the objectives or curriculum of a school, a knowledge of the aptitudes, the goals, and other characteristics of the students is essential. Thus, a study of the student body is necessary to provide background data for assessment in other areas.

Evaluation Problems

As is true with other areas, evaluation problems with respect to the student body arise from the question, "In what ways do good institutions differ from poor ones?" Each answer suggests a dimension along which a measurement might be made if suitable techniques are available.

Size. One obvious characteristic of a student body which shows variation from institution to institution is its size. The important question is whether the number of students enrolled can provide information about the quality of the school or about attitudes toward the school on the part of prospective students and their parents. Certainly some schools purposely keep their enrollments down, and obviously some schools are located in sparsely populated areas and would not be expected to have a large student body. On the other hand, if account of administrative policy and the available supply of students in the community is taken, perhaps some insight into the quality of the school will be obtained from the size of the student body.

There is, of course, the question of how much variation in size of institution will remain when the extraneous variables have been controlled. A perusal of the 1956 College Blue Book indicates, however, that even in this age of expanding college populations a surprising number of schools are not operating at their full capacity. This suggests that, particularly among small private liberal arts colleges, there still exists some competition for students. The use of student-body size as an index of the quality of a school would undoubtedly lose some of its utility with the anticipated great floods of students in the 1960's.

While the general trend of college attendance is upward, some census figures suggest that there will be minor year-to-year fluctuations. To the extent that this occurs, corresponding fluctuations in the student-body size may be indicative of school quantity; one would expect those schools which are well regarded to maintain a fairly constant or uniformly increasing size of student body, while those of poorer reputation would have a student-body size which waxed and waned with the size of the school-age population.

It would appear that indexes which are less subject to the trends in nationwide college enrollment could be derived from a study of the group of persons who apply for admission to a particular institution.

Even if every college and university had more students applying for admission than it could handle, there would still be considerable variation in the ratio of applicants to openings. Care must be taken, however, not to rate low those schools which have accepted their responsibility to make their facilities available to as many as possible, while giving high ratings to those who make no effort to meet the incoming tide. Nevertheless, if appropriate controls are maintained on such things as type of school, community situation, and school policy for regulating the number of its admissions, consideration of the number of applicants can provide a measure for evaluating an institution.

Quality. It is a fact well recognized in athletic circles that to produce winning teams requires talented players as well as good coaching and physical facilities. The same principle applies in academic work. Unless a school can attract its share of good students, it is not likely to maintain a high place among institutions of its kind.

It certainly seems reasonable that the school of a given type which provides the most stimulating environment is likely to attract the most desirable students. Those high school students who have shown outstanding ability and impressive records will, through their advanced study or their teachers, have heard that certain schools are outstanding and will make a special effort to obtain admission to a particular place. On the other hand, the choice of the less able pupils will tend to be more greatly influenced by such other factors as where their friends go, distance from home, football fame, etc. The quality of the student body and the applicant population is, therefore, one factor which must be considered when evaluating institutions of higher education.

In passing, it should be said that the quality of the student body may not be an important dimension for the evaluation of all types of institutions. For example, some community schools may have the avowed purpose of fulfilling the higher educational needs of a restricted geographical area, and may not be interested in attracting outstanding students from other communities. In such a case a really good institution may have only average students. Thus, while the quality of the student body cannot be overlooked when evaluating an institution, the indiscriminate use of this characteristic in all situations is to be avoided.

Achievement of the student body. A third major dimension along which students might be expected to differ is their general achievement during and after their college days. Here one would consider all of those things which could not be said to directly or uniquely reflect the effectiveness of classroom instruction or the quality of curricular offerings. Again, being mindful of the type of institution and making evaluations only in terms of similar schools, studies concerned with the eminence of graduates, the national honors won by the students, the admission of graduates to professional schools, the achievement of students on tests of admission for graduate and professional schools, etc., would provide evidence as to the quality of the school.

The use of achievement without reference to aptitude seems on the surface far more reasonable in a general evaluation than it does when the effects of a specific learning situation are being assessed. Certainly, in some sense that institution is the better which turns out the better product, regardless of how much grinding and polishing has been necessary. However, if one feels that an important function of a university or college is to educate people, then the aptitude and previous achievement of the students must be given some consideration in any evaluation.

Morale of the students. A fourth important characteristic of student bodies which might vary with the quality of an institution is what might be called student morale. Just as in any work situation, morale is an important determiner of output. If conditions are such that the students as a whole are antagonistic toward their immediate surroundings, there is little likelihood that they are getting the greatest possible benefit from their education. It would seem, therefore, that the operating principles of the better institutions would be those which tend to promote the high morale necessary for optimum learning.

Indications of student morale would appear in a variety of ways. The most direct evidence would come from expression of the students themselves. Are they glad they have selected the school they did? Do they feel it is an honor to be a member of the student body? Do they take pride in their campus? Less direct but not less valuable indications of morale could be obtained through a study of the amount of participation in campus activities, the number of group disciplinary problems, the reasons for transfer to another school, the attrition rate, etc.

Personal and demographical characteristics. A final dimension, important for evaluation, along which institutions might differ involves such things as socio-economic level, family and educational background, place of residence, reasons for attending college, educational, vocational, and life goals, etc. While most such personal and demographical characteristics serve primarily as background data to be considered in making evaluations in other areas, some may be used directly. For example, it seems likely that by and large the better schools will have a reputation far more widespread than that of the poorer schools. Therefore, a measure of the extent of reputation might be obtained from a study of the geographical distribution (making adjustment for over-all school size and the school policy with respect to admission of out-of-state students) of the student body and/or applicant population. Similarly, information about the students' motives in attending college might provide direct evidence of the quality of the institution.

Techniques of Evaluation

The techniques which might be used to obtain evaluations along each of the dimensions discussed in the preceding section are not much different from those described in the first three chapters. Nonetheless, illustrations of their use in the study of student body characteristics would seem worth while. Before turning to the separate discussion of the various available techniques for studying the characteristics of a student body, reference should be made to a study by Williams (71) as the best example of the attempt of an institution to describe a portion of its student body comprehensively. In this investigation, through the use of interview, factual information, rating scales, and measures of aptitude, attitude, and adjustment, a carefully selected sample of 100 students was studied intensively with respect to a wide variety of characteristics, including such things as vocational and educational adjustment, recreational and social life, and family environment.

Tabulation of factual information. The tabulation of factual information is of course one of the most important techniques for studying student-body characteristics. The usefulness of the method is obvious with respect to determining size of the student body, applicant population, and the like. With respect to the achievement of students, tabulations of the number of honors students have won, of the number of students successfully completing professional schools, etc., provide valuable indexes. Similarly, counts of the number of transfers and drop-outs, of membership in student organizations, and of attendance at student meetings are important in determining the morale of the student body.

While no study using such factual information specifically for evaluation has been found, two publications have a direct bearing on the topic. In the first, the General Committee for the Cooperative Study of Secondary School Standards has suggested the use of this method and worked out forms on which the information obtained can easily be recorded (81). Although some adaptation would be necessary before these forms could be used in any particular situation, they do provide a basic example of the kinds of things which might be done.

In the second study, McNeeley (42) presents a study of student mortality in which comparisons were made among twenty-five institutions. Although the purpose was not to evaluate the institutions but rather to relate attrition to the changes in behavior and to the personal and background characteristics of the students, the study does show that there is considerable variation from school to school with respect to the variable under consideration.

Personal data sheet. Closely associated with the tabulation of factual information is the recording of information obtained on a personal data sheet. Questions about the student's high school record, family background, vocational plans, reasons for going to college, and reasons for selecting the institution he did, included in registration or application materials, provide valuable data for characterizing both the applicant population and the student body.

In spite of the teacher-to-teacher and school-to-school differences, high school grades still remain the best single predictor of college achievement. This is probably because they include, in addition to an element of aptitude, some combination of the student's motivation and his ability to affect other people favorably, both of which are important determiners of ultimate accomplishment. Whatever the reason, as long as past achievement is predictive of future success, the high school record will remain an important item in the assessment of the quality of college applicants.

In addition to providing supplementary material for judging the quality of the students, the data concerned with family background, vocational plans, reasons for going to college, and reasons for selecting a particular institution yield information useful for determining the reputation of a college or university.

Robertson (60) has investigated the use of the personal data sheet in colleges and universities. He found that 334 different questions were being asked in 78 typical colleges. A few states, such as Minnesota and Ohio, have subsequently worked out a single personal data form for use in all colleges of the state, thus simplifying the work of high school administrators, facilitating transfers, and aiding comparative studies. However, on the whole, the situation has not improved greatly with respect to obtaining comparable data in each school. Such a requirement is necessary if any normative data are to be obtained.

The reliability of the personal data sheet has been studied by a number of authors. Notable among such studies is that of Thisted and Jones (68). Comparing answers to questions which were identical in two different information blanks given two weeks apart, they found that 47 percent of the items had a reliability of .80 or greater, 69 percent of the items had a reliability of .70 or greater, and only 9 percent a reliability of .59 or less. Although reliabilities are for the most part satisfactory, it is surprising to find that seemingly simple items of information are sometimes not answered with a high degree of consistency by students.

Although no studies using data from personnel information sheets have been used directly to evaluate an institution, some studies have made comparisons of two or more schools on the basis of the characteristics of student bodies. The study by Simenson and Geis (64), who compared student bodies of the universities of Oslo and Wisconsin, is noteworthy because of the procedure used. Rather than obtaining a complete census, careful random samples of the student bodies of the two schools were obtained. Thus an accurate comparison could be made at a cost far less than would otherwise have been possible.

Aptitude and achievement tests. The use of achievement tests helps overcome difficulties of evaluating college applicants on the basis of their high school records. These difficulties result from differential grading policies and standards which vary from school to school and from teacher to teacher within the school. The use of such tests also provides an assessment of achievement less likely to be influenced by a halo effect and by the student's personality than are school grades. Thus a better estimate of the extent to which an applicant possesses the knowledge and skills required for college work can be made, and this in turn provides a somewhat different basis for judging the quality of the applicant population.

Aptitude tests provide data for the assessment of applicant population and student body on a somewhat different basis. Here an attempt is made to consider potential ability more or less independent of the effects of a particular school experience. Besides furnishing this direct measure of a particular quality, aptitude tests provide essential control data for making certain comparisons of achievement (for example, in a study of the eminence of former students).

The idea of using the results of aptitude and achievement tests to make comparisons of student bodies from school to school is of course not new. As early as 1927, the Educational Records Bureau published normative data on aptitude and achievement tests taken by freshmen of member colleges. Originally, the colleges were listed and the quartiles obtained by the freshmen at each school indicated. With this information, a college officer could make direct comparisons of the quality of the freshmen at his school with that of the freshmen at other, similar schools. Later, however, the data were presented not for individual colleges but for the following classes of colleges: coeducational, men's, women's, private, public, four-year, junior, and teachers'. In 1940, the regular publication of even this information was discontinued and the readers referred to the special study series of the American Council on Education. Thus, today the college officer who wishes to compare scores made by students in his institution must rely on the "national norms" of such published tests as the ACE, the General Culture Test and the Graduate Record Exams. The many limitations of such national norms as are now available have been discussed elsewhere and need not be repeated here.

In addition to this data on a national basis, a number of states have testing programs from which valuable information for evaluation purposes can be obtained. Segel (63), in describing these programs, lists ten states which do testing specifically for college admissions purposes. In the Minnesota program, for example, the ACE and the Cooperative English Test are given to high school students throughout the state. Every fall the colleges and universities in Minnesota can obtain the scores for all students coming from Minnesota who have been admitted as freshmen to their institutions. Thus information is readily available for making comparisons among institutions with respect to the aptitude and achievement of their new students.

Questionnaires and attitude scales. The administration of questionnaires and attitude scales to members of a student body can provide considerable information not otherwise obtainable. Surveys which ask the students to express directly their feelings toward an institution as a whole as well as toward its various aspects will yield data which, in combination with other material, can result in an excellent index of student morale.

When possible, student attitudes toward other institutions should be included. By asking the students to assume that they could go to any school they wished and having them rank a number of schools in order of preference or choose between institutions presented in pairs, data would be obtained which could be analyzed by one of the multidimensional scaling procedures. This information, combined with knowledge of the characteristics of the schools, would result not only in an indication of which institutions were generally most preferred by students but also in considerable insight into the bases of such preferences.

Perhaps the most extensive study of student-body characteristics using the questionnaire method is that done by Frederiksen and Schrader (22). This investigation involved comparisons of veterans and non-veterans in sixteen institutions of higher learning throughout the country. In addition to test scores, the responses to a questionnaire of some forty-five items were obtained, and direct comparisons of the responses were made among institutions as well as between veterans and non-veterans. Many of the questions were such that valuable information could be obtained which permitted inferences about the quality of the institutions concerned. For example, the percentage of affirmative answers to the question whether the student would go to the same institution if he had a completely free choice ranged from only twenty-eight in one institution to eighty-eight in another. Typical of other questions important for evaluation purposes and having a wide school-to-school variation are those concerned with reasons for going to college, amount of effort exerted in studies, general satisfaction with the education being received, opinion of the quality of instruction and quality of the various student personnel services in the institution.

More recently Pepinsky (55) has described a number of studies of student-body characteristics which have made use of questionnaires and scales. Most interesting for purposes of evaluation are those studies which have revealed discrepancies between student and faculty conceptions of various aspects of the university. For example, one study carried out by Fletcher and Pepinsky indicated that in the College of Engineering the faculty wanted the students to become high-level research workers, while the students themselves were looking forward to operations jobs in managerial or sales work.

Another study mentioned by Pepinsky involved the development by Edwiges de C. Florence of a Student Behavior Description Inventory designed to evaluate the student's goals. Comparisons among institutions of the results obtained through the use of this forced-choice, self-description instrument might well provide valuable information for evaluation.

Evaluation Needs

Because so little seems to have been done by way of evaluation of groups of institutions through a study of student-body characteristics, it is possible to present the evaluation needs by outlining a more or less ideal research program. Such a program serves as a general pattern for evaluation in all areas, and the needs discussed in other chapters represent those items which, when combined with current knowledge, will complete this pattern.

Studies of school characteristics. It has been pointed out in previous sections that meaningful assessment can take place only within groups of schools classified by purpose, size, location, administrative policies, etc. Therefore, the first step in the development of an evaluation system is to survey the schools to determine which characteristics might be used for a meaningful classification. Also included in such a survey should be a number of the items mentioned in the discussion of the tabulation of factual information. Here the purpose would be to show empirical evidence that colleges and universities do vary along the dimensions which have been suggested.

A start in this direction has already been made. Mollenkopf (45) describes a study of school characteristics at the high school level. The purpose of this study is to determine what characteristics of secondary schools are most closely related to the performance of their students on aptitude and achievement examinations. As a start, a 25-item questionnaire was mailed to the principals of a sample of public secondary schools throughout the United States. Thus data was obtained from 844 schools on 32 variables reflecting school, parental, and community characteristics. When the final results of this study are reported they will be of considerable help in the general problem of school evaluation.

A somewhat similar study at the college level has recently been proposed by McConnell (41) and is now in progress. Assuming that the needs of students and society will require an even more diversified higher educational system than now exists, McConnell hopes a survey of school and student characteristics will "...suggest a means by which more effective diversification might be attained without sacrifice of a high level of intellectual accomplishment in liberal and professional education." To achieve these goals, this author suggests that comparisons within various types of institutions with respect to such things as aims, selection of students, scope and depth of educational programs, and organizational structure be made. He also suggests that a number of studies within institutions would be valuable. Among those pertinent to this discussion are the following:

1. A study of the degree of heterogeneity of student bodies through a description of student population in terms of such characteristics as academic aptitude, geographical origin,

socio-economic and cultural background, interests and motives,

2. A study of the relationship between the aims of an institution and these characteristics of the student body, and

3. A study of the relationship between the degree of the heterogeneity of the student body and the diversity of the educational program.

When Dr. McConnell completes this project, an important step will have been taken toward providing information essential to the evaluation of an educational institution.

It must be noted that if studies of school characteristics are to be done reliably, it is essential to standardize still further the demographic terminology and the means of measurement. For example, the mere comparison of enrollments is no simple matter. Should all students be included even though some are part-time? Should they be counted and compared as of a particular date or should an average be taken for the year? Should night-school or summer-session students be included? Similarly, in figuring faculty size should one include administrative officers, personnel workers, librarians? Some standardization on these points has been achieved by the U.S. Office of Education, but more is needed, with particular reference to the purpose of each study.

Development of appropriate measuring devices. A second important phase in the development of an evaluation system is the development and refinement of appropriate measuring instruments. While the development of a number of objective and reliable indices based on factual information is readily accomplished, considerable painstaking research is often required to produce satisfactory measures of many personal characteristics. Yet, if evaluation is to be complete, this must be done.

With respect to aptitude and achievement testing, a number of excellent measures are now available at the college freshman level for many subject-matter areas. At the more advanced level, however, much remains to be done. Except for a few scattered tests for particular professions, achievement at the advanced levels has hardly been touched. Although the Graduate Record Examination and the Miller Analogies Test are widely used general aptitude tests for graduate students, practically no tests of special aptitude in the upper ranges have been published. Some work, however, is currently being done in the development of tests to measure such things as creativity and scientific aptitude. Until more adequate tests of achievement and special aptitude at the upper levels are available, a weak spot in the evaluation of colleges and universities through the study of students will remain.

Even less progress has been made in the development of measures which would provide indications of student morale. Although a number of vocational and academic interest tests have been published, and generalized attitude scales such as that developed by Remmers might be used, such measures for the most part have shown very little relationship to college success. On the other hand, a study recently reported by Frederiksen and Melville (23) does offer some encouragement for the development and use of interest measures when appropriate control variables are used. They found, for example, that interest scores were far more predictive for non-compulsive persons than for compulsive students, and concluded that it is possible to improve the usefulness of many of our tests by discovering subgroups of individuals for which the test is especially effective. This would seem to be particularly true with many of our attitude and personality tests, but much more work needs to be done to verify this suggestion.

While work on the measurement of motivation has lagged, at least three studies in the area have produced results sufficiently encouraging to warrant further exploration. McClelland has developed a projective test of achievement motivation. After preliminary experimental work with the test, Ricciuti (58) concludes that the test "...shows encouraging although not immediately practical validities in predicting high school grades after ability differences are ruled out." Hills (30), dissatisfied with ordinary goal-discrepancy scores as measures of aspiration which might be indicative of motivation, developed a scale to measure an individual's economic, social, academic, and professional ambition. Judging from his preliminary results, this inventory shows considerable promise as a measure of motivation which would be extremely useful in an academic situation. Finally, Dr. Benno G. Fricke, now at the University of Michigan, has developed a test in which the student shows his preference for one statement over another and the strength of his feeling on the matter. Administration to hundreds of freshmen, all with high ability scores, shows that strong achievers tend to have a different configuration of answers from under-achievers. The test, currently being refined through empirical validation, holds promise of enabling college administrators to better assess the non-intellectual -- the personality and motivational -- qualities of prospective students.

One final area in which research in the development of measuring devices is needed before evaluation can be complete is that concerned with estimating the degree of success a person has achieved in his work. Is that man most successful who earns the most money, who is happiest, who is held in highest regard by his colleagues, or who is best known among the general public? If the answer depends upon the line of work, then there is a problem of how to equate the degrees of success across fields. To be sure, there would be some agreement among all persons that certain eminent individuals were successful, and most studies concerned with success have had to be content with a crude dichotomy of eminent vs. non-eminent. In the middle and lower ranges of success, counts of the number of publications, the number of offices held in professional organizations, and the opinions of colleagues have been traditionally used in the professions, and in business the usual index of success has been the salary earned. All of these measures leave much to be desired, and a great deal of original work is needed in the area before anything which could be considered halfway satisfactory as a measure of success is obtained. A possible conclusion is that there are many rather than one criteria of success, some of these criteria perhaps being negatively correlated, and that it will be necessary to study them separately, using appropriate subgroups of the population in question. Until a satisfactory way of quantifying success has been developed, however, evaluation in terms of success of graduates must remain tenuous.

Studies of the relationship among measures. Once relatively adequate measures of a large number of those characteristics of a student body which seem pertinent for evaluation have been developed, it is important to study the relationships among the measures and between each of the measures and other criteria of the quality of the institutions. As each measure is developed it is hoped that it will make some unique contribution to the evaluation. But always an empirical check is necessary, and, more than likely, a number of the measures will be found to be highly related. To avoid redundancy, only one of each group of highly correlated measures should be used in the final analysis.

The comparisons among the methods should go beyond the area of student-body characteristics, for it is possible that measures in one area of evaluation are highly related to those in another. More important, however, is the fact that the various independent measures should all show some consistency in selecting the highest and lowest quality institutions as determined by some over-all or composite evaluation. As in the development of all evaluation batteries, the goal is to obtain a number of measures independent of each other but related to the criterion.

Gathering of normative data. The final phase in the development of an evaluation system involves the gathering of data upon which to base norms for the different tests used. Each measure ultimately selected must be applied to a large sample of institutions of various types. Then the distribution of values obtained from each individual measure or from each composite index must be determined, and standard or percentile scores computed. Finally, a table for converting the raw values to standard measures should be constructed. If data can be gathered for a sufficiently large number of institutions of any given type, then the standardizing process should be repeated for each category, using only the one class of schools each time. To the extent that this can be done the resulting evaluation process will become useful and meaningful.

At present two major difficulties are encountered in gathering adequate normative data. The first of these is a reluctance on the part of school officials to release data on their schools for publication. While in some instances member colleges in such organizations as the College Entrance Examination Board, the Educational Records Bureau, etc., have permitted their data to be included in norms, the results, as illustrated on page 153, are obviously not very useful for making realistic evaluations. The criterion for inclusion of schools in the norms -- that they require certain tests of their applicants -- is not a very meaningful one and, as can be seen from the list of colleges included, leads to a sample far from representative. Also, to be of real value, the entire distribution of test scores, or at the least deciles, is needed, and the colleges should be classified according to curricular type, means of support, and geographical location, as well as by size.

The second obstacle to the publication of adequate normative data has been a lack of uniformity in the tests given by various schools. Educational Testing Service has recently prepared a list of all institutions participating in one or more of its programs. This list shows which tests each college or university is using and represents a

Boys in Liberal Arts, Class of 1958,
Enrolled at Colleges Requiring Both the SAT and Achievement Tests[1]
(mean scores)[2]

College Code Number[3]	SAT-V	SAT-M	English Composition	Social Studies	Average of all Languages	Biology	Chemistry	Physics	Average of Intermediate and Advanced Mathematics
COLLEGES WITH 150 OR MORE SUCH STUDENTS									
A-1	612	618	605	630	606	582	613	593	637
A-2	595	614	586	614	567	582	603	575	611
A-3	582	607	586	596	566	558	597	583	618
A-4	571	615	573	601	521	...	589	571	609
A-5	570	594	571	588	527	...	587	575	605
A-6	558	585	541	606	550	608	597	576	619
A-7	518	536	507	531	503	515	542	523	556
A-8	489	530	493	514	513	530	511	511	529
A-9	479	508	501	495	469	...	474	507	525
A-10	473	489	477	491	415	...	...	...	476
COLLEGES WITH LESS THAN 150 SUCH STUDENTS									
A-11	593	617	587	605	559	...	595	...	626
A-12	574	582	566	577	541	...	593	598	604
A-13	524	562	527	529	...	...	536	548	570
A-14	516	538	513	554	558	...	...	...	545
A-15	486	538	526	544	514	...	539	496	559
A-16	468	477	479	470	398	...	471	...	505
A-17	466	500	472	479	415	...	494	...	492
Average	557	579	555	584	539	553	573	564	585

---

[1] Norms are based on scores received on January, March, or May 1954 tests.

[2] Means were not computed for tests taken by less than 25 students.

[3] College code numbers include the following colleges (not necessarily in order of listing): A-1 through A-10 -- Amherst, Columbia, Georgetown, Harvard, Middlebury, Princeton, Tufts, University of Pennsylvania, Wesleyan, and Yale; A-11 through A-17 -- Brandeis, Bucknell, Carleton, Haverford, Lehigh, Swarthmore, and Villanova.

preliminary survey of the information on file at ETS which might be used for computing normative data.

The results of this ETS survey strongly suggest that some means for obtaining comparable results for a larger number of schools is badly needed. One possibility along this line would be to take a small rigidly chosen sample from a large number of schools rather than using the data obtained on all students from the few schools participating in a given program. A second possibility is to devise a short 15- or 20-minute test which could be given by every school along with its own series of tests. Scores on this common set of items could then be used for equating results of the different tests given in different schools. A final possibility is the use of some application of a procedure, currently under investigation in the research division of Educational Testing Service, which involves the administration of different items, sampled from the same pool, to different students for the purpose of estimating the scores that would be obtained had all the students been given all the items.

Should further investigation show that any of these possibilities is feasible, a great deal will have been accomplished toward obtaining normative data adequate for the evaluation of institutions.

Summary of Evaluation Needs

In the foregoing discussion of evaluation needs a rather ideal research program was outlined. Therefore it would seem well to list here the more important needs specific to the evaluation of institutions by a study of the student body. These are:

1. A greater uniformity in the type of personal data gathered on college students and in the definitions of items on which factual information is to be obtained.

2. Adequate measuring instruments to assess the non-intellectual characteristics of students; for example, measures of motivation, attitudes, appreciations, and morale.
3. Objective indices of such things as the extent of the disciplinary problem found in an institution and the degree of student participation in campus activities.
4. Adequate normative data, obtained through either more extensive sampling or the use of some of the procedures for equating scores on different tests.

PRECEDING PAGE BLANK— NOT FILMED

Summary Table for Chapter VI

Morale of the Students

Evaluation Problems →	Size (Student and Applicant Population)	Quality (Student and Applicant Population)	Achievement of the Student Body	Morale	Personal and Demographical Characteristics
Available or Existing Measures and Techniques for Evaluation	1. Tabulation by class, status, etc.	1. Tests of aptitude and high school level achievement 2. High school academic performance 3. Counts of high school extra-curricular activity 4. Recommendation by others 5. Interviews	1. Advanced achievement tests 2. Tabulation of outside honors won, etc.	1. Questionnaires 2. Attitude of satisfaction scales 3. Rate of attrition 4. Ratings by outside visitors	1. Biographical data sheets
Measurement Needs	1. Uniform reporting forms 2. Uniform definitions of applicants, student body, etc.	1. Measures of motivation, attitude, etc. 2. High-level aptitude tests 3. Uniform definitions	1. Measures of attitudes and appreciations	1. Objective measures of the degree of the disciplinary problem 2. Objective measures of the degree of student participation in campus activities 3. Measures of scholarly interest on the part of students	1. Uniform definitions of family, siblings, residence, etc.
Available Interpretive Data	1. U. S. Office of Education tabulations 2. Census counts	1. Normative data for published tests 2. Intra-state ranking class norms (for some states)	1. Normative data for published tests		1. U. S. Office of Education tabulations 2. Census data
Needed Interpretive Data	1. Normative data on size in relation to capacity	1. More extensive normative data	1. Normative data for all other measures	1. Normative data for all measures	1. Normative data

PRECEDING PAGE BLANK- NOT FILMED

Chapter VII

STUDENT PERSONNEL SERVICES

Introductory Statement

Ever since American schools adopted, shortly after the turn of the century, a philosophy which aimed at the development of the student as a whole rather than of the intellect alone, student personnel services have become an increasingly important part of higher education. Hardly an institution in America today repudiates this philosophy; even those which adhere rather strictly to the intellectual goals of education offer one or more of the variety of services which come under the heading of student personnel. To evaluate a college or university without considering student personnel work would be to overlook an important characteristic of our institutions of higher learning.

The boundaries of personnel services have been specified so broadly as to include all of education (32) and so narrowly as to be synonymous with guidance (76). Williamson (72) presents an adaptation of Meyer's (44) list of student personnel functions which is more reasonable. He includes the following: selection and admission; housing; student orientation; health facilities; keeping of records; counseling; and, finally, employment bureaus. While certainly each of these represents an important aspect of personnel services, some of them are best evaluated by persons in professions other than education. Thus this chapter will not be concerned with such things as health services, housing, or the keeping of records except as they apply specifically to one of the other areas.

Evaluation Problems

Having thus restricted our discussion to those services falling specifically within the realm of professional education, the next step is to specify the dimensions along which institutions can differ with respect to personnel services.

Extent of the services provided. Obviously, institutions may differ in the variety of personnel services offered. One may have only a Dean of Students, upon whom all responsibility for student services devolves; another may have an elaborate system involving many full-time specialists. That an analysis of the provisions a university makes for student personnel work can provide a valuable index of over-all institutional excellence has been demonstrated by Gardner (24), who studied specifically the relationship between the number of personnel services offered and other criteria of educational excellence.

A simple tabulation of the number of different personnel services to be found in an institution is not, however, sufficient. As has been pointed out in earlier chapters, the role of a small private liberal arts school in our educational system may be quite different from that of a technical school or a large state university. Similarly, the students of a school located in a large urban community may differ, in their needs, from those attending a college or a university located in a relatively isolated college town. Thus, as always, any evaluation on the basis of the number and kind of personnel services offered must be related both to the objectives of the institution and to the needs of its particular student body.

Articulation with the academic program. A second way in which institutions may differ with respect to student personnel services is in the extent to which such services are integrated with the regular academic program of the school.

Have the student personnel services developed as a completely separate unit, making little or no attempt to relate their function in the personal growth of the student to his intellectual development, or have the services been specifically designed to aid the instructional faculty in carrying out its objectives and to avoid duplication by providing those opportunities for the all-around development of the students which are not available elsewhere on the campus? Is the faculty kept well informed of the kinds of services provided for the students, and does the faculty make known the kinds of services and information which would help them in the classroom? Is the cooperation such that the teaching faculty frequently and appropriately refers students to the various personnel services, and is information obtained in the personnel office, for example through a freshman testing program, freely given to the instructional staff so that they can make use of it in planning their classroom activities?

The answers to these and similar questions concerning the extent to which the student personnel program is articulated with the other college functions would all provide information essential for the assessment of the student personnel services.

Qualifications of the personnel. A third way in which institutions can differ is in the qualifications of the persons who are responsible for carrying out the personnel functions. Are the counselors, the admissions officers, and the sponsors of student activities persons who have had special training in student personnel work, or are they specialists in other areas, carrying out personnel functions as incidental parts of their main duties? Have the full-time professional persons had adequate training for their specific jobs? These are important factors in determining the quality of university student personnel work. The members of a student personnel staff are just as much faculty of the university as are those whose primary function is the teaching of classes; consequently, all the evaluation problems discussed in the chapter on faculty would be appropriate here. Personnel workers, too, can be evaluated in terms of their training, their experience, their publications, their influence on students, the impression they make on colleagues and supervisors, etc. There is no need to repeat here all of the facets of these problems.

Student performance. Travers (70) has pointed out that guidance is just as much a learning situation for students as is any classroom. The same could be said for most of the other personnel services. Thus, as with the classroom situation, evaluation must be concerned with the effectiveness of this learning situation in producing desirable student behavior.

Because some of the student personnel functions (for example, admission testing and faculty advising) involve all the students, some over-all picture of the students is necessary. Thus an assessment of the personal, social, academic, and vocational adjustment of students and former students may provide information relevant to the efficiency of student personnel services. Obviously, careful delineations must be made if it is desired that the assessments refer to the unique contribution of student personnel services, and not to other aspects of the college program such as curricula, teaching effectiveness, etc. Such things as over-all satisfaction with choice of curriculum, relative freedom from disciplinary problems, and the relative use made of personnel functions might fall in this category.

For those personnel functions, for example vocational guidance and personal counseling, which expose only certain students to specific experiences, the assessment of student behavior should stress change in behavior. After the experience, do students more frequently turn to reliable sources of evidence for making decisions? Are they selecting more attainable goals? Do they get less upset over difficulties? Is there an improvement in grade point average? Do students get along better with others? These and other specific changes should be observable if the personnel services are performing well.

Student and faculty satisfaction. Another aspect of the evaluation of student personnel services which deserves consideration is the extent to which the services provided are judged satisfactory by the students and the faculty. Have unreasonable demands been made of students or faculty members with respect to time or amount of participation? Does a student who has made use of the service feel that he is satisfied? Are faculty members of the opinion that the orientation program was helpful, that records kept are adequate, that the selection and classification of entering students has been of value, etc.? Certainly, if the procedures used in the personnel services antagonize students and faculty, there is little possibility that they fulfill their function well. On the other hand, if the students who make use of the services and the faculty members who observe the results are convinced that the benefits of the services are well worth the time, money, and effort involved, there is some reason to believe that the services are of a reasonable quality.

Satisfaction of the supervisor or outside expert. A final major dimension in the evaluation of student personnel services involves an assessment of the extent to which the procedures used are professionally sound and efficiently applied. Are the procedures based on good experimental evidence? Are they acceptable to experts in the field? Are they the most modern procedures available? Does each routine serve a useful function? Are records accurately kept? Is information so communicated as to provide an adequate basis on which decisions by the student and personnel worker can be made? Does the admission and selection procedure have predictive effectiveness?

For the most part questions of this sort can best be answered by the supervisor of the personnel services in question. However, in some instances, evaluation by an outside expert would be of considerable value. It should go without saying that judgments made either by the outside consultant or the supervisor must be based on careful study of the records and materials used and on systematic observation of the processes involved.

Techniques of Evaluation

As in the other major divisions of college evaluation, a variety of techniques have been used to obtain measurements along the dimensions described in the previous section. Most of these techniques will, by now, be familiar to the reader. Nevertheless, each takes on new advantages and disadvantages as it is applied to a different type of material.

A number of reviews and general discussions (e.g., 6, 70, 73, 74, 76) of work in the evaluation of personnel services have been published. No attempt will be made here to give a critique of or to extend this work. Rather, each of the possible techniques will be discussed, and a description of one or two good studies which have used the method will be given when appropriate.

Check lists. If appropriately used, the check list can provide extremely valuable data for an assessment of college personnel services. It provides an accurate objective inventory of the services provided and of the procedures followed. The study by Gardner mentioned earlier (24) demonstrates the validity of such a procedure in general when used to determine the extent of the provisions. If care is taken to include only those services which are necessary to the objectives of the institutions and the needs of the students, the value would be greatly increased. The general application of check lists to student personnel procedures is on much less solid ground. Assessments made on this basis are valid only when definite knowledge exists about the relative value of the various procedures for achieving specific goals. In our present stage of development such definite information is hardly ever available.

The most elaborate study involving the use of a check list is that described by Gardner. In this investigation information was collected by an observer who visited 57 institutions. In addition to a check list of procedures, data were obtained which provided three criteria of over-all effectiveness of the institutions. As a result, normative data were provided and score cards based on the most useful items were developed. Persons intending to make use of a check list to evaluate colleges would do well to study the general plan of this investigation. Other studies (e.g., 13, 77, 86) have made use of check lists in conjunction with other methods. Especially noteworthy is the work done by Wrenn and Kamm (77), whose investigations covered 14 areas of a personnel program and specifically considered the needs of the students.

Survey of factual information. Closely related to the use of a check list is the tabulation of factual information. Like the check list, such information can provide reliable, objective means by which personnel services can be evaluated. Included in a survey of factual information are such things as tallies of the frequency of the use of services, local validity data for tests used in the admission and classification program, biographical surveys of the experience and training of the personnel workers, and studies of the drop-out rate. Where such measures are related to specific goals of the institution and where it is possible to determine the unique contribution of personnel services, factual information provides one of the best all-around evaluation procedures available.

Evaluation by opinions. Obtaining the opinions of students, faculty or outside experts is another way which may be used to evaluate student personnel services. Generally, these methods involve global judgments made on the basis of various degrees of exposure to the situation being judged. Evidence has indicated that such judgments, even when made by experts, are likely to be unreliable unless the results from a number of judges are pooled. In addition to this problem of reliability, there is the question of the appropriateness of the judges, especially when the judges are faculty members who have had little opportunity to make observations or students who have been served by the personnel services only once or twice. Concerning this latter, for example, Travers has pointed out that persons who have gone to astrologers and fortune tellers are ordinarily quite satisfied that these people have done a good job.

On the other hand, questionnaires are fairly readily obtained or constructed, and if properly used they can provide valuable insight into the way student personnel services are seen by others. Typical of the results of studies which have used opinions of students to evaluate student personnel services are those reported by Paterson and Clark (54). These investigators administered a one-page questionnaire to students who had consulted with faculty advisors. They found that 18 percent of the students considered faculty counseling of "great value," 41 percent "quite helpful," 30 percent of "some assistance," and that over 90 percent said they would urge a brother or sister coming to the university next year to consult a faculty advisor.

Ratings. Quite close to the opinionnaire technique, but more acceptable from a measurement point of view, are rating scales, which list specific observable characteristics and which are used by a rater only after a thorough study of the program. The use of modern recording techniques, particularly in guidance and counseling, makes observation and study of procedures by judges far more feasible today than they have been in the past. A number of rating scales for the evaluation of student personnel services have been constructed. One of the most noteworthy of these was that developed by Brumbaugh and Smith (9) for evaluating personnel work in colleges and universities of the North Central Association. This scale assigned weights to ten different areas of personnel work and included specific performance items. More recently, Wrenn and Kamm (77) have set up a rating system to be used only be a qualified specialist or administrator. This scale covers 14 different functions of a personnel program and lists specific services and provisions necessary to each of these functions. The evaluator rates the degree to which each function is carried out according to the needs of the students.

Not all rating scales, however, need be concerned with the procedures of a personnel program. Rating scales can also be constructed which provide an extremely useful tool for the evaluation of student adjustment. Thus scales of personal, job, and social satisfaction, scales of the extent to which a student feels he has achieved self understanding, and many others can be used to advantage in assessing student personnel programs in terms of their products. The difficulties of rating scales -- halo, leniency, etc., -- have been discussed in earlier chapters and need no repetition here. Often, however, and especially in certain areas of personality, rating scales provide the best means of evaluation currently available.

Achievement measures. One of the more frequently mentioned ways to evaluate personnel services has involved the use of achievement measures. To a great extent, this method seems to stem from the philosophy that the major purpose of a personnel service is to get the student in such a condition that he can profit most from instruction. However, even if this objective is accepted as one major aspect of personnel services, there is considerable danger in using achievement alone as an index of the effectiveness of a student personnel program. First of all, there is no direct evidence, (and among many psychologists there is considerable doubt), that persons "completely adjusted personally and socially" are the ones who make the best scholars. For example, the individual who is completely satisfied with his present status may not be sufficiently motivated to get top grades. Secondly, when the purpose is primarily to raise grades, as is often true in counseling under-achievers, grades can easily be improved by advising the student to take only easy courses. Also, the counseling might be effective in allowing the student to earn the same grades but with less effort. Finally, use of achievement measures may not take account of two important groups, those who fail to follow the advice given and those who have been advised not to continue with academic work.

Mention should be made, since the suggestion is often made in the literature, of the possibility of following up graduates and determining their vocational achievement. It is the writer's feeling that such an attempt is both impractical for most institutions and fraught with too many untenable assumptions about the extent to which success and happiness in adult life are singly determined.

In view of the narrowness of the objectives of personnel services implied and of the logical difficulty in the application of the method, it is felt that the use of student achievement for an over-all evaluation of a personnel program is relatively sterile. On the other hand, grades and achievement tests have a place as criteria for judging the quality of certain specific personnel functions, for example, admissions offices, the treatment of probationary students, and the stimulation of academic interests.

A good example of a study which makes appropriate use of achievement for evaluating personnel work is that recently reported by Klingelhofer (37). This investigation was designed to evaluate the effects of one hour of academic counseling versus the effect of four hours of such counseling, but a comparison was also made between a group who had no formal counseling and the groups who did. In all cases the groups were composed of students on probation, and the criterion for the effectiveness of the advice was the grade point made in the term following the advisement period. No significant difference was found between those receiving but one hour of counseling and those who had four hours, but there was a difference larger than could be attributed to chance between the experimental group and the group which had no counseling.

Experimentation. By far the most sophisticated and probably, from a technical point of view, the most satisfactory way of evaluating student personnel services is by carrying out an experiment specifically designed for this purpose. Travers, who reviewed a number of experiments (70), has quite accurately described the complete experimental method in the context of evaluating student personnel services and has indicated that it includes the following: (a) defining the objectives, (b) specifying the group in whom these objectives are to be achieved, (c) developing instruments to measure the extent to which the goals have been achieved, and (d) carrying out a controlled study of student change.

According to Travers, one of the biggest technical difficulties in carrying out such an experiment is the matching of control groups. After reviewing these studies he lists two frequently used, fallacious ways of handling such an experiment. The first was simply letting those who sought the counseling services be the experimental group, those who didn't, the control group. This procedure biases results in favor of counseling, since those who seek help have, by this act, already demonstrated their desire to improve and ability to make some kind of an adjustment. The second fallacious procedure is to make comparisons between those who follow the advice given them with those who do not. One difficulty here is that those who don't follow such advice may show a basic personality problem which interferes; another is that the advice may be poorer in the case of those who didn't follow it.

Except possibly for the particular criterion used, one of the better experimental attempts to evaluate one aspect of student personnel services is that of Blackwell (5). She evaluated the effectiveness of a guidance program by comparing the increase in grade-point average from pre- to post-guidance sessions of an experimental group with the change in grade point during the same period for a similar group which received no guidance. With 40 subjects in each group and controls on scholastic aptitude and college year, the results indicated that the experimental group improved grades while the control group did not.

Stone (66) reports an experimental evaluation of a program of vocational orientation in which the criterion was the appropriateness of vocational choices at the end of the year as contrasted with similar choices made at the beginning of the year. In this study three groups of freshmen were compared. One group was enrolled in a vocational orientation course; a second was given a test battery and received counseling; and the third was exposed to both the orientation course and to test interpretation through counseling. The results indicated that neither the course nor the counseling alone was nearly as effective as the combination of the two.

Making comparisons of the performance of students in control and experimental groups is not the only form of experimentation which can be used to evaluate the effectiveness of a counseling or other student personnel services, however. Berdie (4), for example, reports an interesting study of the consistency of counselors' judgments with respect to the appropriateness of vocational choice. Twenty case histories, including, in addition to the usual information, an indication of what vocational choice was ultimately made, were compiled and presented to several counselors for review. Each counselor was asked independently whether he thought the choice was appropriate. Also, the original counselor was presented the same material and asked the same questions at a later time. Complete agreement was found in 84 percent of the cases, yielding a tetrachoric correlation of .86 between the original judgment and those of the five other counselors. One hundred percent agreement was found between the two judgments made by the same man from time to time.

While this study could have been much improved had a larger number of cases been used and suggestions for vocational choice made without knowledge of the actual choice, it does suggest a technique which could provide valuable information in assessing the counseling program. If a number of "dummy" case histories were to be carefully set up and a questionnaire asking for specific judgments similar to those a counselor would need to make in handling the cases devised, a form of a "counseling proficiency test" could be set up. A key could be developed by the counseling supervisor or other experts, and, over a period of time, appropriate norms could be gathered. Then an evaluation of at least one part of the counseling process could be made by asking the counselors of the school to study the case histories and make the judgments required on the questionnaire.

Evaluation Needs

A comparison of the evaluation problems described in the first section of this chapter with the discussion of the various techniques in the second section immediately brings into focus some of the important research needs in the evaluation of personnel services. Five of the most urgent of these will be briefly discussed in this section.

Studies of procedures. It has been mentioned that a check list is one of the most valuable instruments for use in evaluating the extent of student personnel services. As has been suggested by a number of authors, before this useful approach can be applied to the evaluation of student personnel services more information is needed about how good the procedures are. Studies are needed which demonstrate the influence of various procedures on other criteria of efficiency and which compare the efficiency of alternative procedures in accomplishing specific purposes. A beginning of this kind of research has been made in some areas. For example, Meehl (43) reviews a number of studies comparing clinical and actuarial methods of making predictions about the success of students in academic work. Even this problem has not yet been completely resolved, and very few if any similar studies have been made involving other aspects of student personnel work.

To be sure, the evaluation of personnel services through a study of the procedures followed will always be an indirect approach. Indirectness does not, however, imply anything about usefulness. When appropriate information is at hand, inventories of procedures will provide a simple, objective and valid way for assessing student personnel services. But first we must know what the good procedures are.

Inter-school norms. A second need in the evaluation of personnel services is one which is also felt in several other areas. As in all instances where no set of generally acceptable absolute standards has been or can be developed, the results of a quantitative evaluation are interpretable only in terms of comparisons with similar measures obtained from like institutions and comparable student groups. With the exception of some early work done by Gardner and the North Central Association, there is little evidence of any attempt to gather normative data across institutions.

The gathering of such data would require that the same measures be applied under similar conditions to a large number of institutions. To the extent that the institutions can be classified according to such characteristics as means of support, major curricular emphasis (e.g., liberal arts vs. technical), size, location, etc., the resulting norms will become more and more meaningful. If an agency outside of the schools were to construct such a set of norms and to provide them to the institutions, confidential information need not be revealed to others.

Until such norms are constructed, interpretation of measures obtained will depend upon subjective standards and beliefs about what other institutions are doing, which may show considerable fluctuation from time to time and from school to school. One danger in this is the tendency for some institutions, ignorant of precisely what others are doing, to remain complacent in the belief that they have modern, up-to-date personnel services which are equal to those of the best institutions. When norms have been built, such complacency will rest, at least, on factual information.

Needs analysis. A third blind spot in the evaluation of student personnel services arises from the failure of persons concerned with the problem to consider adequately the needs of the students. Though it is obvious that the sole function of a personnel service is to satisfy some student need, very few evaluations have made any attempt to assess student need and to relate the services provided to these needs.

To adequately determine student needs, a great deal more information about the student is required than is conventionally obtained. It would be important to know, for example, what some of the home conditions are, why the student chose the school and curriculum he did, how much vocational information he actually had, what his hopes in terms of college life are, what he intends to do when he leaves school, and what his life-long ambitions tend to be. This, in addition to factual information about the housing conditions, community recreational facilities, etc., would provide considerable insight into student needs.

Once the student needs have been specified on the basis of such information, evaluation of student personnel services can be made according to the degree to which such needs are fulfilled. This would provide a much more acceptable criterion than is now available.

Adjustment measures. A fourth need is for adequate measures of student adjustment. If student personnel services are to be evaluated in terms of the changes produced in students, then reliable and objective assessments of the student's characteristics before and after he makes use of the services are required. Possibly because of the lack of good measuring instruments in other areas, most studies involving student change have been concerned with changes in grades. It would seem exceedingly important, however, to consider other aspects of the student's development as well.

Although some scales of curricular and vocational satisfaction have been developed, none seems to have been widely accepted. Similarly, measures of adjustment are yet to be developed which are sufficiently concerned with the kinds of changes which might result from the use of personnel services to be of any great value in assessing such services.

The development, then, of acceptable inventories, rating scales, and other measures of satisfaction with curriculum, with social life in the school, with recreational activities, etc.; measures of the extent to which a student has achieved his desired goals, of self understanding, of the extent to which decisions are based on evidence rather than on pressures, irrational feelings, or vague conceptions, measures of the degree of emotional upset over minor problems, and many other specific kinds of adjustment -- such development is work of considerable importance to the future evaluation of student personnel services. Once such measures have been devised, use of the ultimate criterion of the value of student personnel services -- the extent to which the services produce appropriate changes in behavior -- becomes a possibility.

Comparisons of evaluation techniques. The final evaluation need in personnel services is one which was implied but not specifically mentioned in connection with studies of procedures. This is the need for a comparison of a large number of approaches to evaluation made on the basis of a single set of data. It results from the unfortunate fact that the rationally most acceptable procedures for evaluating personnel services are, practically, the most difficult to apply. If, then, it can be shown that the more simple, easily obtainable measures are highly related to the ideal criterion, the former can be used for evaluation purposes with a great deal more confidence. Thus, for example, if the results of a simple check list of procedures or of the opinions of students agree with an evaluation of institutions based on a controlled experiment using student change, then there is no need to go through the elaborate experimental procedure except periodically to recheck the validity of the other measures. A number of comparative studies of this sort would, therefore, point out the number of readily obtainable measures which have a sufficient degree of validity for continued use in evaluating student personnel services.

Summary of Evaluation Needs

In summary, then, the main evaluation needs in the area of student personnel services are these:

1. Studies which demonstrate the influence of the presence or absence of various personnel procedures on the different criteria available, and which compare the efficiency of alternative procedures in accomplishing specific purposes.
2. Normative data on new and existing measures, and indices which will permit comparisons among institutions similarly classified according to such characteristics as means of support, major curricular emphasis, size, and location.

3. Studies which analyze the needs of college students through an intensive study of their personal and background characteristics and through a survey of conditions in the college community.
4. Adequate measures of student adjustment in non-academic areas.
5. Studies which make comparisons among the various possible approaches to evaluation.

Summary Table for Chapter VII

Evaluation of the Student Personnel Services

Evaluatio Problems →	Extent of the Services Provided	Articulation with the Academic Program	Qualifications of the Personnel	Student Performance	Student Satisfaction	Satisfaction of the Supervisor
Available or Existing Measures and Techniques for Evaluation	1. Check lists	1. Check lists 2. Opinions 3. Achievement tests 4. Records of grade-point changes after counseling	1. Frequency counts of number of degrees 2. Distribution of the number of years experience 3. Ratings of credentials 4. Ratings by colleagues 5. Counts of membership in and certification by professional organizations 6. Counts of the number of publication ratings by outside experts	1. Standardized tests of adjustment 2. Control group experiments using measures before and after counseling 3. Follow-up of counselees 4. Attrition studies	1. Questionnaires 2. Rating scales	1. Questionnaires 2. Rating scales 3. Systematic observation of materials used 4. Systematic observation process (e.g., tape recordings, films)
Measurement Needs	1. Uniform definitions of the services	1. Adequate analysis of needs 2. Adequate measures of student adjustment in non-academic areas	1. Records of effectiveness as a counseling intern	1. New and improved measures of adjustment and attitudes 2. Improved design of the studies 3. Objective indices based upon systematic records of behavior (e.g., number of dates, hours of study, etc.)	1. Objective indices of satisfaction	
Available Interpretive Data	1. Inter-college norms--see Gardner (24)			1. Normative data for published tests		
Needed Interpretive Data	1. More recent and more extensive normative data	1. Normative data 2. Comparison studies	1. Normative data for all measures	1. More extensive normative data on published tests 2. Normative data for all new measures	1. Normative data	1. Normative data 2. Relationship between specific procedures and student changes 3. Relationship between specific materials and student changes

PRECEDING PAGE BLANK-- NOT FILMED

Chapter VIII

COMPREHENSIVE EVALUATIONS

Introductory Statement

The preceding chapters have dealt with the evaluation of particular facets of an institution's activities, e.g., the curriculum, classroom instruction, and student personnel services. A good deal has been done in recent years to develop techniques for evaluating these divisions of an institution's program; the results have contributed to the improvement of the quality of higher education. For many purposes this analytical type of evaluation is essential.

In addition to the evaluation of the several phases of an institution's program, it is important to know something about the over-all effect of a college or university -- how well it functions as an institution. Indeed, it may be that in the final analysis the evaluation of the total program is the most important. The alumnus of College X is not the product of the curriculum alone, the classroom, the counseling office, the extra-curriculum, or any other single phase of the institution's program; rather, the alumnus is the product of all of the influences which have been brought to bear upon him during his undergraduate days. This total effect, or impact, should be included in any attempt to arrive at an evaluation of the quality of a particular institution.

Evaluation Problems

Performance of graduates. As mentioned previously, the evaluation criterion having the greatest face validity is that of measuring the success of an institution's graduates in terms of occupational success, personal adjustment, effectiveness as a citizen, creative thinking, and, perhaps, prestige. The use of this type of criterion would make the evaluation of colleges and universities comparable to that of product evaluation in business and industry. A study of the success of an institution's graduates, therefore, has a good deal of appeal. While this type of measurement should be investigated, it must be recognized that the success, adjustment, or effectiveness of college graduates is influenced by many factors, of which the college experience is only one. The selection of students with a familial background of wealth and prestige might, for example, be a more important factor in later success than what happened to these students in college. And, following graduation, the type of job taken, the community of residence, the economic situation, geographical area, and home and family affairs all have a profound effect upon the accomplishments of a particular individual. Some factors can be controlled statistically, but because it is difficult to assess the magnitude of these factors it is very difficult to design a study in which statistical controls are really adequate. Nevertheless, the performance of graduates has been used in evaluation studies, and when the results are carefully interpreted, such studies can contribute materially to an over-all program of institutional evaluation. To be sure, such studies cannot point out directly the reason for a very good or very poor over-all evaluation, but they can stimulate a search for the factors which are important.

Scholarly production of the faculty. Scholarly production, as suggested earlier, could be used in the evaluation of a college or university faculty. It would also seem to be a means of assessing the over-all quality of an institution. Many authorities in higher education would contend that on the graduate level, particularly, quality of teaching and scholarly production are closely correlated. In recent years, foundations, government, business and industry have made substantial grants to educational institutions to encourage research and scholarship. The grants which have been made for non-applied research and scholarship are particularly significant. Largely as a result of these grants, the amount of research and scholarship on the college campus has increased notably since the war. This makes it all the more appropriate that scholarly production should be one dimension which can be used to evaluate the quality of an institution of higher learning.

Scholarly atmosphere. Closely associated with scholarly production, but not identical with it, is the factor of "scholarly atmosphere." On some college campuses intellectual work ranks high in prestige, not being overshadowed by extra-curricular activities and service functions. Unfortunately there are also some campuses where the reverse is true. One of the striking characteristics of some institutions is that departments work closely together, crossing subject-matter boundaries freely in order to work on particular research problems or to develop interdepartmental programs of study. There are also institutions in which departmental boundaries are rigid and where a minimum of cross-fertilization of ideas occurs. Freedom of teaching and scholarly pursuits is often called the sine qua non of higher learning. Nearly all colleges and universities subscribe to this philosophy of education, yet there are decided institutional differences. On some campuses there is an atmosphere of conservatism, of concern, and even of fear. In contrast there are campuses where there is not only complete fearlessness about the presentation of new ideas, but a constant prodding to be courageous, imaginative, and venturesome. The latter point of view is admirably expressed in the words of Redfield:

> 'May it [the University of Chicago] strive always to enlighten, knowing it will frequently be misunderstood. May it fall neither into complaisance nor into popularity. May it ever put creativity above conformity and intellectual courage

> above security. In the eternal questioning that is its productive life may it often include a re-examination of its own nature and purposes. May it be forever discontented that its discontent lead to new excellence." (56)

Whether or not a college has the type of atmosphere described by Redfield makes a great difference in its effectiveness as an educational institution. There is no doubt that "atmosphere" is a dimension or a scale along which institutions will show considerable variation.

Meeting the needs of society. In the second chapter it was stated that one dimension for the evaluation of objectives should be that of "appropriateness," that is, an institution's objectives should be judged in part in terms of the adequacy with which the stated objectives meet the needs of the community served by the institution. This statement could be applied equally well to the institution as a whole. Institutions of various types -- liberal arts colleges, teachers colleges, junior colleges, universities, and colleges or units within universities -- perform unique functions. The curriculum, the faculty, student personnel services, and other activities presumably are designed to enable the institution to play its proper role in the scheme of things. A teachers college which fails to turn out graduates who enter the teaching profession is not meeting the needs of society as effectively as another teachers college which does produce teachers. A liberal arts college which turns out semi-specialists rather than liberally educated young men and women (these two functions are, of course, not necessarily incompatible) is not a good liberal arts college. In other words, each institution has a role to play if it is to meet effectively the needs of society. All facets of the institution's program must be designed to achieve this end. The degree to which the institution succeeds in playing its role would seem to be one criterion which could be applied to determine its total quality.

Impact on higher education. It is a well known fact that some institutions have had much more impact on higher education than have the vast majority of colleges and universities. Colleges such as Antioch, Bennington, St. John's, and Sarah Lawrence, to mention only a few, have developed novel approaches to the undergraduate education of students. This element of novelty or creative imagination is something quite apart from the factor of the quality apparent in the curricula, facilities or personnel services of colleges in general. Universities such as Minnesota with its General College, Iowa with its advanced degrees in the creative arts, Columbia and Chicago with their programs in general education, and Wisconsin with its experimental college are pioneers in new educational developments. Many universities have very good faculties, curricula, and research facilities, but only a few have led the way in bold new ventures and have consequently made a noteworthy impression on the whole field of higher education. To be sure, an institution should not be rated high simply because it has started one or more new programs. It is a fact in education as in other fields that many hunches and new ventures do not work out successfully. Nevertheless, if an institution never tries anything new, it is unlikely to make much of an impression upon education. Institutions do differ in their vigor and imagination, and in general it would seem that those of the highest quality have contributed the most new developments.*

---

*There probably are a few institutions which have done their teaching, research, and service jobs so well that they have achieved great distinction without engaging in many new ventures. However, these would appear to be the exception rather than the rule -- and it is possible that they might have been even greater had they been more venturesome.

Techniques of Evaluation

Ratings by experts. The quality of educational institutions is assessed most frequently through some sort of personal judgment. When parents and children are making plans for the childrens' college education, some form of rating is probably applied to the institutions under consideration. When young faculty members have a choice of institutions at which to work, some sort of subjective evaluation of the institutions is likely to be made. State and regional accrediting agencies when attempting to determine whether a given institution should be placed on the accredited list usually arrive at a final judgment based in considerable part on opinions of the over-all quality of the institution. The judgmental process, assisted by check lists and rating scales, is probably the most favored method for measuring the general excellence of colleges and universities.

In the North Central study frequently mentioned in the preceding chapters, Haggerty (48 , pp. 11-14) describes the comprehensive or over-all evaluations of the 60 institutions included in the early stages of the study. Three judges, Presidents Gage and Zook and Dr. Reeves, were asked to rank, independently, the sixty institutions in order of general merit, giving a rank of 1 to the best institution and a rank of 60 to the poorest. The correlations between the independent rankings were surprisingly high. The correlations, computed by the Spearman method, were as follows: Gage-Zook, .93, Gage-Reeves, .87; Reeves-Zook, .86. It seems clear from these results that well-informed experts in the field can agree quite closely in evaluating the general excellence of educational institutions. To be sure, these high correlations do not necessarily guarantee high validity of rating. It is possible that all of the judges were influenced by such extraneous factors as beauty of the campus, appearance of the college catalog, size of endowment, the reputation of the college and, perhaps, by some general, unverbalized stereotype of what constitutes "goodness" or "quality" rather than by real educational excellence. All that can be said is that three competent judges were able to agree among themselves as to the over-all quality of a group of colleges and universities.

Since the early experience with over-all ratings in the North Central study was encouraging, it was decided to incorporate this type of rating in Criterion VIII, the criterion which was used for determining the excellence of the 57 institutions finally included in the study. Accordingly, three over-all ratings were made and included in Criterion VIII. One was a combined rating by Reeves and Gregg, based on the general excellence of the institutions, and made after the institutions had been visited. A second rating was made by Gardner, also on the general merit of the institutions, and one by Haggerty, based upon four separate judgments, one of which was "institutional tone," a second, "attainment of educational objectives," a third "order in which one would recommend the institutions to young men and women, other things such as cost being equal" and fourth "a rank order from the standpoint of prestige, as determined by history, traditions of merit, reputation of faculty, endowment and other sources of stability."

An interesting approach to the rating of the administrative reputation of departments in a college of liberal arts is described by Hemphill (29). Among other things, an attempt was made to identify the best- and worst-administered departments in the college, using a nominating technique. Each faculty member participating in the study was asked to name the five departments which had the reputation of being the best led or best administered departments in the college (a total of 22) and also the five departments which had the reputation of being least well led or least well administered. With these nominations in hand it was possible to compute "reputation scores" for the 22 departments in the college. It would be difficult to establish the validity of these reputation scores, but it was found that they did have a high degree of reliability. While the technique would seem to function best within a single institution, it might find some application in the larger field of institutional evaluation.

The use of ratings in the over-all evaluation of institutions is characterized by the strengths and weaknesses inherent in any rating process. The ratings will be as good or as bad as the quality of the judges, the information available to them, the freedom from bias, prejudice, and other extraneous factors, and the clarity of definition of the factors to be included in the rating. As stated on several earlier occasions, rating requires that the human mind function as a measuring instrument. This is fraught with many difficulties, and yet, from many points of view, particularly because of the complexity of educational institutions, it would have to be said that the final judge of the quality of a college or university must be the expert mind. Perhaps for this reason the North Central Association relied heavily upon human judgment in making the evaluation of the educational institutions included in its study and has continued to use human judgment extensively in its accrediting process. Despite all the measuring techniques which have been developed in recent years, it seems inconceivable that a comprehensive study of general excellence of colleges and universities would be carried out without giving a heavy weight to human judgments.

Follow-up studies of graduates. One of the most attractive suggestions frequently made concerning college and university evaluation is that the quality of an institution should be judged solely by the product which it turns out, namely, its alumni and former students. In effect, this suggestion means that institutions of higher education should follow the success (broadly defined) of their graduates, and evaluate their educational endeavors in terms of the achievements of those who have been exposed to their influences. The use of follow-up studies has been discussed in earlier chapters. Since it is so difficult to isolate the relative influence of classroom teaching, library facilities, curriculum organization, and personnel services, it would seem that the follow-up study of graduates would be peculiarly adapted to the over-all evaluation of educational institutions. A number of studies have been carried out, designed especially to provide information about the relative success of various types of institutions in producing certain types of graduates.

One of the most recent and widely publicized follow-up studies of college graduates is that of Knapp and Goodrich (39). This study, undertaken by the Wesleyan Committee on the Education of Scientists, was directed to discovering the origins of scientists, particularly the undergraduate colleges in which they studied and the types of influences to which they were subjected while in college. The criterion group of scientists consisted of those listed in the third (1921) and seventh (1944) editions of American Men of Science who held the Ph.D. degree and those non-Ph.D.'s who were starred in this publication. The science fields represented were mathematics, physics, chemistry, astronomy, geology, biology, and psychology. For each scientist, the name of the college granting the baccalaureate degree and the date of that degree were ascertained. All data were then tabulated by science field and institution. To equate for differences in size of institutions, a "productivity index" was computed by dividing the number of male scientists by the number of male graduates during the period 1924-34 and multiplying this number by 1000. This index gives the number of scientists produced per 1000 male graduates. It was thus possible to compute a productivity index for each of 490 institutions.

After the productivity indexes had been computed, statistical studies were made of the institutions included in the study, questionnaires were sent to a group of scientists, and actual visits were made to twenty-two institutions -- those seeming to be representative of the group, or possessing some unique features. The institutions with the highest productivity indices were found to be fairly small in size, of the liberal arts type, and located in the Middle and Far West.

It was also noted that all of the Catholic institutions were found in the least productive 10 percent of the institutions studied.

In addition to the general study of 490 institutions, more intensive studies were made of 154 colleges and 84 large universities. In the case of the colleges, eighteen factors were studied in order to explain, if possible, the individual differences in productivity indices. The factor showing the highest correlation with the index was geographic location; other factors which appeared in the multiple correlation were endowment per student, minimum student cost, and average size of graduating class. Specifically, the Middle West and Far West colleges had the highest productivity indexes; there was also a tendency for those with higher educational costs, higher endowment per student, and larger average size graduating class to have slightly higher productivity indexes, but these factors were not of any great significance. The correlation of geographic location with productivity indexes was .388; the addition of the other three factors raised the multiple correlation to .445.

One factor which proved to be very significant, but which could not be studied in all institutions, was that of the mental ability of the students. American Council on Education Psychological Test scores were available for fifty schools, and for this group it was found that, in general, the higher the average score in the test the greater the productivity of the scientists.

A study similar to the college study was made of 84 universities. For this group it was found that the most productive institutions were those which had average, or slightly above average, costs of attendance, were located in the Middle and Far West, and had the higher faculty-student ratios. No data were available on the relative quality of the students in the various universities included in the study. Two factors, geographical location and cost of attendance, were common to both the college and university groups.

Following the completion of the general statistical studies, visits were made to the 22 institutions which seemed to have unique educational programs and which had very good records in the production of scientists. The visitor to each of these institutions made ratings of such factors as the library, laboratory facilities, faculty research, use of student assistants, individual study programs, grading standards, requirements for a major, and "degree of social interaction." Since only a small group of institutions was visited, no very definite conclusions could be drawn from the findings. It did appear, however, that severity of major requirements and quality of departmental *esprit de corps* were associated with the productivity index.

A second study, by Knapp and Greenbaum (40), is also of interest. This study shows the relationship between college of undergraduate study and success in advanced study as measured by the earning of a Ph.D. degree since 1948 in one of the 25 universities sampled, winning a university fellowship or scholarship since 1946 from one of these 25 universities, receiving a private fellowship since 1946 from one of the 9 private foundations sampled, or being awarded a government fellowship since 1946 from the United States Public Health Service, the Atomic Energy Commission, or the Fulbright program. A further condition for inclusion in the study was that those included have graduated from college in or since 1946. A total of 7,000 individuals met these criteria.

Several indices of productivity (of scholars) were computed; of these the All Male Index seems the most significant because of the large number of male graduates in the colleges studied. The All Male Index was computed for 377 institutions. It indicates the number of scholars (as defined above) per thousand graduates during the period 1946-1951. Other indices computed were the All Female Index, "Partial" Male Indices (for the humanities, social sciences, and sciences separately). Analyses were made by the four types of award received and by the classification of the undergraduate institution, namely, liberal arts colleges, universities, technical institutions, women's colleges, and a residual group of miscellaneous institutions.

The chief findings of the study are as follows: (1) institutions of high cost seem to produce the largest number of scholars; (2) the New England and North Central regions are more productive than other regions, science students appearing to predominate in the North Central region and humanities students in the New England region; and (3) among both liberal arts colleges and universities the non-denominational institutions have the highest productivity indices.

One of the striking facts about the latter study is the contrast with the findings of the earlier study of scientists. Especially significant is the fact that the high cost, New England institutions were found to be very productive of scholars in this study, whereas in the study of scientists, the medium cost, Middle Western and Far Western institutions were found to be the most productive. The chief factors explaining this change seem to be the G.I. Bill and the new scholarship policy of the costly Eastern institutions. As a result, more students of the type who previously went to lower-cost institutions were able to attend the more costly colleges and universities. It should also be remembered that the criterion was different in the two studies, the earlier study being made of established scientists while the latter was of young scholars, many of whom were still in graduate school when included in the study. The results of these two follow-up studies dramatically illustrate how important is careful consideration of the design of a study and how striking is the difference brought about by changes in criterion measures.

The studies by Knapp and Greenbaum and by Knapp and Goodrich represent interesting and provocative approaches to the investigation of the origins of American scientists and younger scholars. These studies are described here in some detail because they make use of techniques which might well receive wider application -- especially if certain improvements were made in the application of the techniques and in the interpretation of the results. For example, in the Knapp and Goodrich study it appears that the liberal arts college is a better producer of scientists than is the university; in actual fact, this may or may not be the case. First, the typical university has a large number of professional and semi-professional undergraduate colleges and curricula (e.g., commerce, engineering, pharmacy, nursing, agriculture) whose graduates should not be considered potential scholars. By contrast, the graduates of liberal arts colleges are, to a greater extent, potential material for the graduate colleges. Second, many university liberal arts colleges permit students to earn a B.A. degree by counting one year of professional college study in a college of law, medicine, dentistry, etc., toward that degree. While these students earn the B.A. degree, they are not comparable, in terms of their potential interest in graduate study, to students majoring in other departments. Third, as pointed out by Knapp and Goodrich, the orientation of the students themselves is a factor of major importance. While the undergraduate college or university will influence the development of students very considerably, their basic orientation probably is not changed a great deal. In other words, it would appear that students select their colleges because they have certain interests; it may not be so much a matter of having certain interests because they attended a particular college. If studies such as that of Knapp and Greenbaum are to be useful to individual colleges, there would need to be adequate statistical control over the predictor variables associated with academic and vocational success.

The studies by Knapp and Goodrich and Knapp and Greenbaum were restricted to students who went on to graduate colleges. One of the more comprehensive follow-up studies of college graduates in general was that carried out under the auspices of Time and reported by Havemann and West (26 , pp. 171-185). The chapter entitled "Princeton versus Podunk" is especially relevant. It was found that the graduates of Ivy League schools, notably Princeton, Yale, and

Harvard, earn substantially higher salaries than graduates of other types of colleges. Even when allowance is made for differences in the socio-economic status of the students' parents and in the types of fields entered, the differences in average salaries still seemed to persist. The authors suggest that differences in intellectual abilities (because of selective admission policies) and motivation are factors which may serve to explain the "success" differences. In other words, it would seem that the higher salaries earned by Ivy League graduates should not always, if ever, be attributed to the superiority of their educational programs.

The chief defect in the Knapp-Goodrich, Knapp-Greenbaum, and Havemann-West studies is the failure to control the variable of intellectual ability. (Ideally, factors such as familial background, motivation, and professional goals should also be controlled, but intellectual ability would seem to be the single most important factor. If all the colleges and universities studied in these investigations had administered the same measure of scholastic aptitude, e.g., the American Council on Education Psychological Examination, the performance of graduates of various types of schools could have been compared on a more equitable basis. Also, if the Knapp-Goodrich and Knapp-Greenbaum studies had been expanded to include students going on to all types of professional schools, the results would have been of greater value in the rating of the educational quality of the undergraduate colleges and universities. The follow-up studies here reviewed are definitely on the right track, but they would have to be sharpened up a good bit in order to make them useful in the evaluation of various types of institutions of higher learning.

A follow-up study which definitely did take general mental ability into account has been reported by Wolfle and Smith (75). The population studied consisted of high school graduates, of whom some went to college while others did not. Salary earned fifteen to twenty years following graduation from high school was the criterion used. The study makes no comparison between graduates of various types of colleges, but it does illustrate very well the types of comparisons which can and should be made if the factor of scholastic aptitude is taken into consideration.

As mentioned in Chapter III, one of the earliest and still one of the best follow-up studies of college students is that of Pace (50). The population studied in this instance was restricted to former students of the University of Minnesota. Many of the same techniques of measurement were used, however, in the study by Havemann and West. In other words, the techniques developed and used by Pace in his study are useful in making comparison between colleges as well as in studying the performance of students who attended one particular institution.

Achievement measures. Along with the follow-up study of college graduates, it has been suggested that colleges and universities should be evaluated in part by the performance of their students on standardized achievement tests. The Pennsylvania Study by Learned and Wood, the North Central Study frequently mentioned in this report, the results obtained by the use of various Cooperative Tests and the Graduate Record Examination, all reveal wide institutional differences in the achievements of students as measured by standardized achievement tests. As one examines these findings one cannot help but feel that these test scores do reflect, in part, differences in the educational quality of institutions in which the students are or have been enrolled. But before placing too much stock in test scores as measures of the excellence of educational institutions, the reader should remember two very important facts; namely, that the differences in student aptitudes are probably just as great as they are in achievement, and that the curricular requirements of colleges show wide variations. Both these factors have been discussed several times in earlier chapters.

If, however, one could get agreement on the validity of an achievement test battery such as the Graduate Record Examination, and if all colleges would agree to give the same or comparable tests of scholastic aptitude, it would seem that achievement measures in relation to aptitude would contribute a great deal to over-all college and university evaluation. But until these problems are met it will be necessary to accept with strong reservations the results of achievement tests as measures of the educational quality of various types of institutions.

Evaluation Needs

The evaluation of the general excellence of colleges and universities is probably equal in importance to that of the evaluation of particular facets of an institution's program. The distinguishing feature of over-all evaluation is that it is the sort of evaluation which cannot be referred to some specific aspect of the college -- for example, the quality of its graduates. Regional accrediting agencies are particularly interested in this type of evaluation, whereas professional accrediting bodies are more concerned with individual departments, divisions, or schools of an institution. A particular college or university may wish to know where it stands among other colleges or universities of its type. Having this general information, it can proceed to ascertain its particular strengths and weaknesses. There are, then, many reasons why the comprehensive evaluation of institutions is important. What is needed by way of techniques and interpretive data to improve the validity of this type of evaluation?

Better rating methods. The North Central Association of Secondary Schools and Colleges and other accrediting agencies make extensive use of the method of ratings by experts. In order to facilitate this rating procedure, the North Central Association has developed the Revised Manual of Accrediting, 1941. Detailed directions are given for carrying out the rating process, weighted scores are provided for the assigned ratings, percentile scores are given for the several variables, and, finally, the results are entered on what is called an Institutional Pattern Map. In general, this procedure would seem to be sound. The errors or problems encountered in this procedure have been mentioned several times in this report. Any rating is certain to be influenced by the quality of the judge. Since the Manual is used by different judges, individual differences in rating are introduced. The weighted scores assigned to various ratings are largely subjective and, of course, the final judgment is and has to be subjective. Nevertheless, it seems unlikely that this general procedure can or should be abandoned in the forseeable future. Care in the selection of judges, greater precision in the writing of the directions, more refined weights, and truly representative percentile norms would go far toward the improvement of this rating procedure. Basically the method appears sound; what it needs is refinement.

Better follow-up studies. The chief defects in the follow-up studies completed to date have been mentioned in the preceding pages of this chapter; failure to recognize differences in scholastic aptitude of students, failure to control such factors as the backgrounds, attitudes, and plans of students, inadequate sampling of various groups of students, and the use of questionable criterion measures. The use of salary earned as the sole criterion of success and the failure to recognize differences in student aptitudes are the most serious criticisms of the Havemann-West study conducted under the auspices of Time. From the standpoint of sampling, this study would seem to be very good. The Knapp-Goodrich and Knapp-Greenbaum studies are good examples of intensive studies of one type of graduate, the one going on to graduate study. The criteria of success used in these studies seem quite reasonable; the sampling, too, seems satisfactory.

The ideal type of follow-up study would seem to be one modelled after that of Havemann and West, but on a very large scale and using success in graduate or professional school (with appropriate correction for differences in the scholastic ability of students and other predictors of success) as well as salary earned within occupational groups as criteria of success. The results of such an ideal study should still be interpreted with caution, but they would be much more appropriate than those reported to date.

Better achievement measures. The need for better measures of educational achievement has been stressed repeatedly in this report. In the over-all evaluation of institutions the need is particularly great for comprehensive test batteries, such as the Graduate Record Examination and Cooperative Survey Tests, which really measure all the objectives of instruction. Test batteries of this kind have worked out best on the elementary and secondary levels (e.g., the Iowa Tests of Basic Skills and the Iowa Tests of General Educational Development). The problem of building these tests on the college level is formidable. But until this is done, achievement measures will have serious limitations as techniques for the evaluation of colleges and universities.

Summary of Evaluation Needs

The major improvements in the techniques and interpretive data needed for better over-all evaluation of colleges and universities are presented in the accompanying table. The ideal type of over-all evaluation of colleges and universities would seem to be a combination of ratings, follow-up studies, and achievement measures. Ratings by highly trained, competent judges should be made of those variables which do not lend themselves to objective measurement, e.g., the impact of the institution upon the educational world, the scholarly atmosphere of the institution, and the extent to which it meets the needs of society. Large-scale, controlled follow-up studies should be made of the graduates of various types of institutions in terms of their success in graduate work, in professional schools, and in some of the major occupational fields. The academic proficiency level of undergraduate students should be measured by a Graduate Record Examination type of test battery at the end of the senior year, all students having previously been given the same measure of scholastic aptitude. With the data at hand from ratings, achievement test batteries, and follow-up studies it should be possible to arrive at an over-all evaluation of institutions superior to anything which has thus far been reported in the professional literature of education. The chief defect in the Knapp-Goodrich, Knapp-Greenbaum, and Havemann-West studies is the failure to control the variable of intellectual ability. If all the colleges and universities studied in these investigations had administered the same measure of scholastic aptitude (e.g., the American Council on Education Psychological Examination), the performance of graduates of various types of schools could have been compared on a more equitable basis. Also, if the studies by Knapp and Goodrich and by Knapp and Greenbaum had been expanded to include students going on into all types of professional schools, the results would have been of greater value in the rating of the educational quality of the undergraduate colleges and universities. The follow-up studies here reviewed are definitely on the right track, but they would have to be revised a good deal in order to make them useful in the evaluation of various types of institutions of higher learning.

PRECEDING PAGE BLANK- NOT FILMED

Summary Table for Chapter VIII

Over-All Evaluation

Evaluation Problems →	Performance of Graduates	Scholarly Production of the Faculty	Scholarly Atmosphere	Meeting the Needs of Society	Impact on Higher Education
Available or Existing Measures and Techniques for Evaluation	1. Ratings 2. Biographical data 3. Test scores (e.g., GRE) 4. Diary keeping (of behavior)	1. Ratings by experts in the field 2. Tabulations of number of publications 3. Counts of number of honors to faculty members	1. Ratings 2. Questionnaires	1. Ratings	1. Ratings 2. Tabulation of the number of new programs tried
Measurement Needs	1. Development of criteria of success	1. Objective indices of the amount of faculty time spent in research and other scholastic production 2. Tabulation of the proportion of funds spent for research	1. Tabulation of the number of cultural activities 2. Index of amount of funds spent for outside lecturers 3. Objective measures of the degree of academic freedom	1. Objective measuring devices	1. Additional objective measuring devices
Available Interpretive Data	1. Normative data on published tests 2. Productivity indices for various institutions--see Knapp and Greenbaum (40)				
Needed Interpretive Data	1. More extensive normative data for all measures	1. Normative data for all measures	1. Normative data for all measures	1. Normative data for all measures	1. Normative data for all measures

PRECEDING PAGE BLANK- NOT FILMED

Chapter IX

RECOMMENDATIONS

Each of the preceding chapters contains a summary, in tabular form, of the measurement techniques and instruments and of the interpretive data which we feel are needed in order for colleges to do a better job of making self evaluations. These statements of measurement needs may be thought of as detailed recommendations which grew out of this survey.

It may be desirable to present, in addition to the tabular summaries, an overview of these measurement needs. The direct needs may be thought of as falling into three general categories: (1) testing devices for measuring a variety of things important in college evaluation, for example, those hoped-for outcomes of instruction which are not well covered by existing tests; (2) more adequate normative data for existing tests, and similar interpretive data for any new measuring instruments which may be developed; and (3) methods for improving the process of making judgments by such groups as evaluation committees. All of these needs are directly concerned with improving the process of evaluation. In addition, there are some categories of evaluation needs which are concerned with certain questions underlying the needs for measuring instruments: (1) what specifically are good and poor practices in teaching, counseling, and other college functions? and (2) what are the skills, knowledges, and abilities which might be taught in college and which are related to success after college? The first three types of needs are concerned directly with how to improve the process of evaluation; the last two are concerned with the basic problem of what to measure. These five types of needs will be discussed in the following sections.

Direct Evaluation Needs

Development of adequate tests. For a number of aspects of evaluation the use of tests has been mentioned as an appropriate method. However, for many important variables tests do not exist, and for others existing tests are not entirely satisfactory. For example, individual faculty members are likely to be reluctant that evaluation of their work be based on available achievement tests, on the grounds that the tests do not adequately represent some of the really important things being taught. Such attitudes cannot be ignored, partly because the grounds for rejecting the tests may be valid and partly because, for a self-evaluation to be effective in producing eventual improvement, it is vital to have the full cooperation and acceptance of the methods by all the people concerned.

Thus, one of the important needs is for achievement tests which reflect more adequately those important instructional objectives which have so far largely eluded the producers of standardized tests. Such new tests would serve to supplement existing tests which do cover certain aspects of the problem quite adequately.

Some of these objectives lie in the area of attitudes and appreciations. The sociologist may hope his students will acquire greater sympathy for the problems of underprivileged peoples, for example, or the teacher of English may wish his students to acquire a taste for good literature. Published achievement tests may do a good job in providing measures of amount of factual knowledge learned or of degree of skill acquired in understanding and interpreting written material, but techniques for measuring the less tangible but no less important objectives such as those illustrated above are largely undeveloped.

Even in the cognitive area there may be important skills not now measured by even the best of standardized tests. In order to make large-scale testing feasible, we have tended to adopt objective types of test items, such as the multiple-choice, almost exclusively, with the result that there is some validity to the charge that the examinee doesn't have to think of anything -- he only has to recognize or evaluate what is put in front of him. Skills such as ability to organize thoughts and data for logical presentation, or ability to make a wise selection of information for a certain purpose or to think of reasonable hypotheses to account for observed phenomena tend to go unmeasured by standardized tests. These deficiencies in standardized tests largely result from demands for efficiency and economy in testing programs. A research program aimed at the development of testing instruments for measuring these more complex cognitive skills should, by directing attention toward these important intellectual skills, have a very wholesome effect on teaching.

These new testing devices of course need not be paper-pencil tests at all. The imaginative use of situational or performance tests or systematic observation of spontaneous behavior in certain situations might make possible the inclusion of a number of important educational objectives in the evaluation system.

In addition to the need for more adequate achievement tests, there is need for objective measures of various other attributes of a college, as described in earlier chapters. In the summary tables have been mentioned, for example, need for an instrument to determine what are the faculty members' conceptions of the college's objectives; measures of the usefulness, to the student choosing a college, of catalog statements of objectives; a device for objectively measuring articulation between school and college; scales for measuring faculty attitude toward teaching; and objective measures of student participation in various kinds of campus activities.

Useful interpretive data. The most obvious need for norms is in the area of aptitude and achievement tests. Norms are needed because most aptitude and achievement tests, unlike physical measures, do not have a scale with units of equal size and with a known zero point. Such tests yield numbers which permit the rank-ordering of examinees but do not permit comparison with any absolute standards. Exceptions might be tests like tests of typing ability, where one can get absolute measures, such as number of words per minute, which have meaning without reference to a set of norms. Most tests, however, are composed of items of unknown difficulty (in the absolute sense), and the only way one can make the scores meaningful is by comparing candidates, or groups of candidates, with one another.

In order that a college, then, may tell how able its students are, it must compare them with other groups, such as students in other colleges. Most test publishers make available normative data of some sort to facilitate this process. The most usual practice is to publish a single table of norms which may be labelled "national" norms. Rarely is it justified to accept the norms as really national, however; they are most likely to represent the colleges that happened to buy the test offered by the publisher, or the colleges that the publisher could persuade to participate in a norming study. Rarely is the collection of colleges well enough described to be very helpful in interpreting the results.

Another possibility would be to make available as norms the distributions of scores obtained at various colleges which are identified by name. Then the dean could find out how his engineering students compared with those of MIT, or how his junior-college freshmen compared with the liberal arts freshmen at Oberlin. Unfortunately, test data of this sort are considered confidential, and college officials are very reluctant to allow them to be published. This attitude has been successfully combated in some instances; for example, Robert North, formerly of the state testing bureau in Kentucky, persuaded a number of colleges to permit publication of norms for use by high school counselors. But, by and large, colleges -- especially those who think they might suffer by comparison with others -- are not willing to have their score distributions published.

Further consideration of the uses of norms as a means for judging the relative standing of an institution soon leads one to the conclusion that neither the one set of "national" norms nor norms for identified colleges would be an adequate solution to the problem. On the basis of national norms a few colleges would consistently be above the 95th percentile on most measures, while others would consistently be below the 5th percentile. The colleges above the 95th percentile would probably be much more interested in comparing themselves with a small group of colleges like themselves than with the national norms. Norms for single institutions, unless very extensive, are likely to suffer from lack of data at critical places.

Norms for subgroups of institutions defined by region, type, source of support, size, etc., would perhaps be the best solution. Such norms would make it possible for the college to compare itself with institutions which it considers like itself. A college at the 5th percentile on national norms might find that it compares very favorably with colleges of similar size and objectives in its region, even though it is very low by national standards.

All that has been said about aptitude and achievement test scores for students applies equally well to most of the other measures which are likely to be obtained in a college evaluation program. After finding the average annual number of publications for the liberal arts faculty, one needs to be able to compare it with averages for other colleges in order to judge the relative productivity of one's faculty. And again, in order for the comparison to be meaningful, one needs norms for institutions classified by size, type, region, etc.

In the case of achievement tests given to students at the end of one or more years of college training, norms differentiated on still another basis, in addition to size, type, and region, would be extremely helpful from a college-evaluation point of view. The additional factor to be taken into account is the initial ability of the students, as measured by suitable tests. The purpose of a college must involve producing some change in student level of competence, and in order to evaluate such change it is necessary to take into account initial level of ability. One way to take account of initial ability would be through differentiated norms. Separate norms on an achievement test in mathematics, for example, might be presented for colleges whose mean scores on an entrance test of mathematics fell in certain ranges.

The recently completed ETS study at the secondary school level, mentioned in Chapter VI, has identified some of the principal dimensions on which schools differ and has found how performance on a variety of aptitude and achievement tests varies with school characteristics. The study indicates that by far the best predictor of mean score (by school) on achievement tests is mean score on aptitude tests. On the basis of this school characteristics study it would seem feasible at the high school level to publish norms differentiated both by type of school and by ability level. Differentiated norms would be equally feasible at the college level.

Methods for producing the needed normative materials will now be discussed.

One possibility is to conduct a large-scale data-collecting study expressly for the purpose of obtaining norms. Such a study might involve the following steps:

1. The first step would be to determine specifically the variables on which norms are to be obtained. The variables would include not only the aptitude and achievement measures and the biographical and attitude data judged appropriate for students, but also the many variables involving faculty, objectives, curriculum, and personnel services discussed in earlier chapters.

2. Next it would be necessary to choose or to develop actual workable measuring instruments to represent the variables chosen. It would be necessary to face the question of how much developmental work to try to do before proceeding with the data-collection phase. One extreme view would be to proceed with whatever instruments happen to be available now; the other would be to wait until a much improved battery of tests is available.

3. The third phase would be to choose the institutions to participate in the study and to persuade them to cooperate. The choice of institutions would have to give adequate representation to all the types, sizes, regions, etc., which would be involved in the differentiated norms. In order to represent each subgroup of institutions adequately, a fairly large number of institutions would be required.

4. Next, the tests and other instruments would have to be administered, taking care that reasonably standard conditions hold.

5. Finally would come the scoring and tabulating and the preparation of normative materials.

Such a study would undoubtedly be costly, but it would provide a wealth of materials which would be extremely valuable to colleges interested in self-evaluation. Such a direct attack on the problem would be most satisfactory if it could be successfully carried off.

Another method, much less expensive, could be employed for the more limited purpose of providing norms of student performance on aptitude and achievement tests; this would involve the use of a short "link" test, or perhaps several such tests, for the purpose of establishing relationships among the great variety of existing tests. Very extensive norms exist, for rather ill-defined groups of colleges, on many tests. The College Entrance Examination Board very carefully scales scores on its examinations for "College Board Colleges." The American Council on Education Psychological Examination is used extensively by another group of colleges which overlaps somewhat with the College Board. The Ohio State University Psychological Examination has good norms -- for colleges in the state of Ohio. The norms for the graduate program tests -- for admission to medical schools, law schools, business schools, etc. -- are normed on applicants, or, in the case of the Graduate Record Examination's institutional testing program, on seniors in the colleges which happen to participate. If some of the scores on all these diverse tests could be related to a common scale, a great deal of information which could be used for comparative purposes would become available.

The notion, then, is merely to arrange for the experimental administration of a short ability test to all the candidates taking the tests in their regular programs. Then it would be possible to relate OSU, ACPE, College Board test scores, etc., to the scale of the short link test. Thus would be provided a means for comparing students in colleges even though different tests are used. The system could easily be kept up to date when revisions of tests are made merely by an administration of the revised test along with the link test to establish the relationship between the tests. The accuracy with which such comparisons can be made will depend upon the amount of correlation of the link test with each of the other tests. Also, problems related to the age or educational status at the time of administration of the test would have to be solved. Such a study should be much easier to set in motion than the one previously described because it would involve much less disruption of the college's regular testing program.

It is apparent that a third possible kind of norming study would be a combination of the best features of the two described above. Such a study might involve the link test method of norming for a variety of aptitude and perhaps achievement tests, plus separate arrangements for obtaining other types of test and questionnaire data.

There are other possible methods for improving the sampling and making the whole process more efficient. For example, it is possible to get normative data without testing entire populations in the selected schools. A larger number of schools could easily be included with comparatively few people per school, provided those to be tested are specifically designated in advance in a manner which minimizes the likelihood of obtaining biased samples. Another possibility is to use a new technique being studied by Lord which involves administering different sets of similar items to the various students in the sample. In other words, by taking advantage of some of the newer techniques of sampling it might be possible to get data for a large-scale and comprehensive norming study at moderate cost.

Improvement in the process of group judgment. In a number of the chapters of this report, the primary method of evaluation available involves a process of judgment, usually by a committee. It is inevitable that in the final analysis much of college evaluation consists in individual or group judgment, since questions of value are involved. Even in areas where standardized tests can be used, the prior problem exists of deciding on the nature of the items to be included in the test. The fundamental reasons for a college to exist are based on value judgments, and some of the areas for evaluation, such as the objectives of the institution, are concerned with these very basic questions of value. A very important need, then, has to do with the improvement of the process of making judgments by groups such as evaluation committees.

The much-relied-upon committee technique of solving problems is presumably based upon the assumption that a group of individuals sharing in the deliberations on a problem will produce better solutions than individuals working alone. But there is a good deal of reason to fear that committee decisions are often of poorer instead of better quality than would be produced by a capable individual. An aggressive member of the committee may persuade the group to adopt an inferior point of view, while the shy individual with a better idea is not heard. Because of the number of people involved, no one feels the full force of responsibility for the action, and hence important aspects of the problem do not receive full attention. The discussion atmosphere may not lend itself to the acquisition of information and the appropriate use of that information in solving the problem. Most participants in committee work can recall instances where a dramatic but trivial item of information crops up, is remembered, and exerts undue influence on the outcome of the discussion.

The problem of group judgment of course extends far beyond the area of college evaluation, and considerable research on group processes is in progress. In the military the concept of staff action has been influential in determining the pattern of problem solving, and the Air Force in particular has supported research on the process of group decision-making. Under the heading of group dynamics other research is going on. The continuance of such research may eventually, it is hoped, result in better understanding of how small group processes may be controlled to advantage and how information may be fed to the group to maximize the probability of sound decisions.

Since values are so strongly involved in the evaluation problems of concern to college faculties, it is clear that more is involved than mere logical or syllogistic problem solving or the application of any existing methodology. The existence of a sound philosophy and sound attitudes must be assumed as basic to good judgment in areas involving value judgments. One would hesitate to recommend an experimental approach to the study of what constitutes a sound philosophy. But research aimed at study of small-group discussion and problem solving might eventually tell us something about how best to organize, instruct, and handle a group in order to enhance the likelihood of its making sound evaluative judgments.

Need for Background Information

What constitutes good practice? When one undertakes to construct a rating form for an observer to fill out on the basis of his observation of a counseling session, for example, he is faced with the problem of what specific items of behavior to include on the form. He may write down with some confidence items concerned with greeting the client in a friendly manner and speaking distinctly enough to be understood. But one soon realizes that there are distinct limitations to our knowledge of what constitutes good counseling practice. One has only to look at the literature to realize that there is disagreement among the experts with regard not only to methodology but also to the underlying theory of counseling. According to some points of view, for example, it is good to present the client with factual information about his performance on scholastic aptitude tests, while from other points of view such information should rarely if ever be given to a client. Until more is known about what constitutes good practice in counseling, the development of effective evaluation methods is obviously limited.

It is easy to see how evaluation in any of the other areas is similarly limited by lack of information as to what constitutes sound practice. Sometimes the lack of information is inherent in the nature of the problem because the problem is purely a question of value and not amenable to direct empirical study. But in many other instances the value judgment has been clearly settled, and the problem is the experimental one of determining which of several possible methods best produces the desired outcome. The answer will not always be simple; the answer may be that for one kind of student Method A is better while for another Method B is better. Or the answer may be even more complicated; the best method may depend not only on the kind of student but also the kind of instructor, the kind of subject matter, and many other conditions. The point is, however, that if we know the good from the poor procedures, the task of evaluation is possible; we have only to develop methods for finding out to what extent the good methods are used in appropriate circumstances.

A great deal of research needs to be done, then, on the problem of what constitutes good practice in such areas as teaching, student personnel services, college admissions procedures, and selection and promotion of faculty. Much work is of course going on, but the problems are legion; acceleration of such research would appear to be highly desirable. Evaluation cannot become objective until we know what are the desirable practices to look for.

What abilities are needed? Just as we must know what are the desirable practices in performing the functions of a college, so must we know what are the desirable attributes of students and graduates if we are to evaluate objectively. Again, some of the problems are intrinsically problems of value. For example, although having neurotic students on the campus may result in annoying problems for the faculty and administration, whether a college can rightfully deny an applicant the opportunity of a college education merely because of evidence that he is neurotic is in large part a question of values. On the other hand, many problems are amenable to empirical study; for example, it can be determined to what extent, if any, the neurotic student is handicapped in his ability to perform the academic work required by the institution.

There are at least two aspects of the question of what abilities are needed. One is the problem of what abilities are desirable in the students admitted to various programs of instruction; the other is the problem of what abilities are required for success in each of the many activities for which students are being trained. The first has many implications for admission and guidance programs, and the second has implications particularly for matters of curriculum and instruction.

With regard to qualities of students, much work has already been done on the relation of intellectual abilities to academic grades, and personality characteristics have also been under investigation. In most of these studies the criterion has been grade, based for the most part on teacher-made tests or standardized achievement tests, both of which are likely to be subject to the weaknesses mentioned in Chapter V. With criteria which reflect more adequately certain desired objectives, perhaps we would have a better chance of fairly evaluating certain kinds of student abilities such as ingenuity, imagination, or creativeness. Perhaps the personality picture of the potential good student would change; some evidence now available suggests that the shy, compulsive, poorly-adjusted student is likely to get the best grades, other things being equal; but with grades that more adequately recognize a student's ability to see beyond the names and dates to the broader picture of what is being presented perhaps different personality characteristics of the good student will be found. More work is needed to find out the characteristics which are desirable in a student. These characteristics will undoubtedly vary somewhat, depending upon the academic program involved. Knowledge of these characteristics will better enable us to evaluate the student body.

Less work has been done on the abilities required for success in law or medical school, in the practice of law or medicine, or in being a good citizen in the community. Here we are concerned primarily with achievement rather than with aptitude. It is the familiar problem of articulation -- in this case, the articulation of the graduate programs. Knowledge of what abilities and skills are required for success in the graduate school should help to determine the undergraduate curriculum. Knowledge of the abilities required for success after graduate training should help to determine the graduate curriculum. Again, only with background information of this sort do we know what to evaluate, and an evaluation program cannot fully succeed unless it evaluates the right things.

Summary

In broad summary, the survey has brought out three direct needs for evaluation materials: (1) the need for evaluation instruments, such as achievement tests which measure important educational objectives not well measured by presently available tests; (2) better normative data to aid in interpretation of data from available tests and new normative data for new instruments where appropriate; and (3) better understanding of how best to organize, instruct, and handle a small group such as an evaluation committee in order to increase the likelihood of its making sound evaluative judgments. In addition, the survey has shown the need for information which will throw light on the problem of what to look for in the process of evaluating. Information is needed as to (1) what constitute good practices in certain aspects of the educative process such as teaching and counseling, and as to (2) what are the abilities needed for success in each of various academic programs and for success in various occupational and community activities after completion of the educational career.

PRECEDING PAGE BLANK- NOT FILMED

REFERENCES

1. Anderson, G. L. "The Reactions of Minnesota College Freshmen to Their First Year of College Experience." University of Minnesota Studies in Higher Education. Minneapolis: University of Minnesota Press, 1941.

2. Beaumont, H. "A Suggested Method for Measuring the Effectiveness of Teaching Introductory Courses," Journal of Educational Psychology, 1938, 29, 607-612.

3. Bethel, L. L. "Vocational Education." In The Public Junior College, the Fifty-Fifth Yearbook of the National Society for the Study of Education: Part I. Chicago: University of Chicago Press, 1956.

4. Berdie, R. "Judgments in Counseling," Educational and Psychological Measurement, 1944, 4, 35-55.

5. Blackwell, E. "An Evaluation of the Immediate Effectiveness of the Testing and Guidance Bureau of the University of Texas," Journal of Educational Research, 1946, 40, 302-308.

6. Blaesser, W. "The College Administrator Evaluates Student Personnel Work," Educational and Psychological Measurement, 1949, 9, 412-428.

7. Bloom, B. S. (ed.). Taxonomy of Educational Objectives. New York: Longmans, Green & Co., 1954.

8. Bond, A. D. "An Experiment in the Teaching of Genetics, with Special Reference to the Objectives of General Education." Teachers College Contributions to Education No. 797. New York: Columbia University Press, 1940.

9. Brumbaugh, A., and Smith, L. "A Point Scale for Evaluating Personnel Work in Institutions of Higher Learning," Religious Education, 1932, 27, 230-235.

10. Chamberlin, D., Chamberlin, E., Drough, N., and Scott, W. Did They Succeed in College? New York: Harper & Bros., 1942.

11. Clark, K. E. "The APA Study of Psychologists," American Psychologist, 1954, 9, 117-120.

12. ___________. Personal communication.

13. Cowley, W. "A Technique for Making a Student Personnel Survey," Personnel Journal, 1931, 10, 17-26.

14. Dressel, P., and Mayhew, L. General Education: Explorations in Evaluation. Washington, D. C.: American Council on Education, 1954.

15. Dyer, H. S., and King, R. G. College Board Scores: Their Use and Interpretation. New York: College Entrance Examination Board, 1955.

16. Ebel, R. L., and Stuit, D. B. A Study of Relationship Between the Educational Backgrounds of Students and Their Performance in Comprehensive Examinations Designed to Measure the Outcomes of General Education. Iowa City: State University of Iowa, 1955.

17. Eckert, R. E. "Evaluation in General Education." In General Education, the Fifty-First Yearbook of the National Society for the Study of Education: Part I. Chicago: University of Chicago Press, 1952.

18. ___________. Outcomes of General Education; an Appraisal of the General College Program. Minneapolis: University of Minnesota Press, 1943.

19. Feil, M. A Study of Leadership in a Liberal Arts College. Columbus, O.: The Ohio State University, 1950.

20. Flanagan, J. C. "The Critical-Requirements Approach to Educational Objectives," School and Society, 1950, 71, 321-324.

21. Flesher, M. A., and Pressey, S. L. "War-Time Accelerates Ten Years After," Journal of Educational Psychology, 1955, 46, 228-238.

22. Frederiksen, N., and Schrader, W. B. Adjustment to College. Princeton, N. J.: Educational Testing Service, 1951.

23. ____________, and Melville, S. D. "Differential Predictability in the Use of Test Scores," Educational and Psychological Measurement, 1954, 14, 647-656.

24. Gardner, D. Student Personnel Service. Chicago: University of Chicago Press, 1936.

25. Guthrie, E. R. "The Evaluation of Teaching," Educational Record, 1949, 30, 109-115.

26. Havemann, E., and West, P. W. They Went to College. New York: Harcourt, Brace and Co., 1952.

27. Heaton, K. L., and Koopman, G. R. A College Curriculum Based on Functional Needs of Students. Chicago: University of Chicago Press, 1936.

28. Hemphill, J. K. "Description of Group Characteristics," Invitational Conference on Testing Problems. Proceedings, 1954. Princeton, N. J.: Educational Testing Service, 1955.

29. ____________. "Leadership Behavior Associated with the Administrative Reputation of College Departments," Journal of Educational Psychology, 1955, 46, 385-401.

30. Hills, J. P. "The Measurement of Levels of Aspiration," Journal of Social Psychology, 1955, 39, 221-229.

31. Hood, P. D., Halpin, A. W., et al. "Crew Member Agreement on RB-47 Crew Operation Procedures," Technical Report. Columbus, O.: The Ohio State University, 1955. (Restricted.)

32. Hopkins, L. "Personnel Procedure in Education," Supplement to Educational Record, 1926, No. 3. Washington, D. C.: American Council on Education, 1926.

33. Johnson, G. B., Jr. "An Evaluation Instrument for the Analysis of Teacher Effectiveness," Journal of Experimental Education, 1955, 23, 331-344.

34. Johnson, R. I. (ed.). Explorations in General Education; the Experiences of Stephens College. New York: Harper & Bros., 1947.

35. Jones, E. S., and Ortner, G. K. "College Credit by Examination," University of Buffalo Studies, Vol. 21, No. 3. Buffalo, N. Y.: University of Buffalo, 1954.

36. Kelly, E. L., and Fiske, D. W. The Prediction of Performance in Clinical Psychology. Ann Arbor: University of Michigan Press, 1951.

37. Klingelhofer, E. L. "The Relationship of Academic Advisement to the Scholastic Performance of Failing College Students," Journal of Counseling Psychology, 1954, 1, 125-131.

38. Koos, L. V., and Crawford, C. C. "College Aims Past and Present," School and Society, 1921, 14, 499-509.

39. Knapp, R. H., and Goodrich, H. B. Origins of American Scientists. Chicago: University of Chicago Press, 1952.

40. __________ and Greenbaum, J. J. The Younger American Scholar: His Collegiate Origins. Chicago: University of Chicago Press, 1953.

41. McConnell, T. R. "The Diversification of American Higher Education," mimeographed proposal of a research study, University of California at Berkeley, 1955.

42. McNeeley, J. H. "College Student Mortality," United States Office of Education Bulletin, 1937, No. 11. Washington, D. C.: Government Printing Office, 1937.

43. Meehl, P. Clinical vs. Statistical Prediction. Minneapolis: University of Minnesota Press, 1954.

44. Meyers, G. "The Nature and Scope of Personnel Work," Harvard Educational Record, 1938, 8, 82-93.

45. Mollenkopf, W. "A Study of Secondary School Characteristics as Related to Test Scores," Research Bulletin 56-6, a multilithed report. Princeton, N. J.: Educational Testing Service, 1956.

46. Morsh, J. E., and Wilder, E. W. "Identifying the Effective Instructors: A Review of the Quantitative Studies 1900-1952," Air Force Personnel and Training Research Center Bulletin 54-44, San Antonio, Texas: Air Force Base, 1954.

47. Newburn, H. K. "The New Program in Liberal Arts," University of Iowa Publications, New Series No. 1350. Iowa City: University of Iowa Press, 1945.

48. North Central Association of Colleges and Secondary Schools, Committee on Revision of Standards for Higher Institutions. Evaluation of Higher Institutions, a series of monographs. Vol. 2, "The Faculty," by M. E. Haggerty. Chicago: University of Chicago Press, 1937.

49. __________ Vol. 3, "The Educational Program," ibid.

50. Pace, C. R. "Follow-Up Studies of College Graduates," Growing Points in Educational Research; Official Report, 1949. Washington, D. C.: American Educational Research Association, 1949.

51. _________. "A Report on the Experiences, Activities and Opinions of 2,500 Syracuse University Alumni," Syracuse University Bulletin, 1950. Syracuse, N. Y.: Syracuse University Press, 1950.

52. _________. "They Went to College." University of Minnesota Studies of General Education. Minneapolis: University of Minnesota Press, 1941.

53. _________. "University-Wide Studies in Evaluation of General Education." In Dressel, P. L. (ed.) Evaluation in General Education. Dubuque, Iowa: Wm. C. Brown and Co., 1954.

54. Paterson, D. G., and Clark, K. E. "Students' Judgments of Counseling," Journal of Higher Education, 1943, 34, 140-142.

55. Pepinsky, H. B. "Productivity in the University," presidential address to the American College Personnel Association, Washington, D. C., March 26, 1956.

56. Redfield, R. "The Dangerous Duty of the University," School and Society, 1951, 74, 161-165.

57. Remmers, H. H. (ed.) "Studies in Attitudes: A Contribution to Social-Psychological Research Methods," Purdue University Studies in Higher Education. Lafayette, Ind.: Purdue University Press, 1934.

58. Ricciuti, H. N. "The Prediction of Academic Grades with a Projective Test of Achievement Motivation: I. Initial Validation Study," Research Bulletin 54-17, a multilithed report. Princeton, N. J.: Educational Testing Service, 1954.

59. Riley, J. W., Jr., Ryan, B. F., and Lifschitz, M. The Student Looks at His Teacher. New Brunswick, N. J.: Rutgers University Press, 1950.

60. Robertson, D. A. "Personnel Methods in College," Educational Record, 1927, 8, 310-321.

61. Roe, A. "A Psychological Study of Eminent Biologists," Psychological Monographs, 1951, 65, 68 pp.

62. Russel, J. H., and Burns, N. "Faculty Status in Member Colleges and Universities of the North Central Association of Colleges and Secondary Schools, 1945-6," North Central Association Quarterly, 1948, 22, 397-413.

63. Segel, D. "State Testing Programs and Services," United States Department of Health, Education, and Welfare Circular No. 447. Washington, D. C.: Government Printing Office, 1955.

64. Simensen, W., and Geis, G. "A Cross-Cultural Study of University Students," Journal of Higher Education, 1955, 26, 21-24.

65. Smith, H. The Purposes of Higher Education. New York: Harper & Bros., 1955.

66. Stone, C. H. "Are Vocational Orientation Courses Worth Their Salt?" Educational and Psychological Measurement, 1948, 8, 161-182.

67. Strothman, F. W. The Graduate School Today and Tomorrow; Reflections for the Profession's Consideration. New York: Fund for the Advancement of Education, 1955.

68. Thisted, M., and Jones, L. "A Critical Analysis of the Personal Information Blank in Use at the State University of Iowa," School and Society, 1931, 33, 540-544.

69. Timmons, W. M. "Decisions and Attitudes as Outcomes of the Discussion of a Social Problem," Teachers College Contributions to Education No. 777. New York: Columbia University Press, 1939.

70. Travers, R. "A Critical Review of Techniques for Evaluating Guidance," Educational and Psychological Measurement, 1949, 9, 211-225.

71. Williams, C. T., These We Teach: A Study of General College Students. Minneapolis: University of Minnesota Press, 1943.

72. Williamson, E. How to Counsel Students. New York: McGraw-Hill Book Co., 1939.

73. ____________, and Bordin, S. "The Evaluation of Vocational and Educational Counseling," Educational and Psychological Measurement, 1941, 1, 5-24.

74. ____________, and Darley, J. Student Personnel Work. New York: McGraw-Hill Book Co., 1937.

75. Wolfle, D., and Smith, J. G. "The Occupational Value of Education for Superior High-School Graduates," Journal of Higher Education, 1956, 27, 201-212, 232.

76. Wrenn, C. "The Evaluation of Student Personnel Work: A Critique of the Guidance Movement," School and Society, 1939, 52, 404-414.

77. ________, and Kamm, R. "A Procedure for Evaluating a Student Personnel Program," ibid., 1948, 47, 266-269.

78. ________. "General Counseling Procedures," in Monroe, W. (ed.). *Encyclopedia of Educational Research*. New York: The Macmillan Co., 1946.

79. "Bridging the Gap Between School and College," *Evaluative Report No. I*. New York: Fund for the Advancement of Education, 1953.

80. "A Design for General Education," *American Council on Education Studies, Series I, No. 18*. Washington, D. C.: American Council on Education, 1944.

81. *Evaluative Criteria, 1950: A Cooperative Study of Secondary School Standards*. Washington, D. C.: American Council on Education, 1950.

82. *General Education in School and College; a Committee Report*. Cambridge, Mass.: Harvard University Press, 1953.

83. *North Central Association of Colleges and Secondary Schools: Revised Manual of Accrediting*. Chicago: University of Chicago Press, 1941.

84. *Requirements for the Bachelor of Arts Degree in the Liberal Arts Colleges in State Universities of the Mississippi Valley*. Iowa City: State University of Iowa, 1955.

85. *A Re-Study of the Needs of California in Higher Education*. Sacramento, California: State Board of Education, 1955.

86. *Western Personnel Institute 1948 Conference Report*. Pasadena, California: Western Personnel Institute, 1948.